The term "nucleic acid" refers to deoxyribonucleotides or ribonucleotides and polymers thereof in either single- or double-stranded form. Unless specifically limited, the term encompasses nucleic acids containing known analogues of natural nucleotides which have similar binding properties as the reference nucleic acid and are metabolized in a manner similar to naturally occurring nucleotides. Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses conservatively modified variants thereof (e.g. degenerate codon substitutions) and complementary sequences and as well as the sequence explicitly indicated. Specifically, degenerate codon substitutions may be achieved by generating sequences in which the third position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues (Batzer et al. (1991) Nucleic Acid Res. 19: 5081; Ohtsuka et al. (1985) J Biol. Chem. 260: 2605-2608; Cassol et al. (1992) Rossolini et al. (1994) Mol. Cell. Probes 8: 91-98). The term nucleic acid is used interchangeably with gene, cDNA, and mRNA encoded by a gene.

"Nucleic acid derived from a gene" refers to a nucleic acid for whose synthesis the gene, or a subsequence thereof, has ultimately served as a template. Thus, an mRNA, a cDNA reverse transcribed from an mRNA, an RNA transcribed from that cDNA, a DNA amplified from the cDNA, an RNA transcribed from the amplified DNA, etc., are all derived from the gene and detection of such derived products is indicative of the presence and/or abundance of the original gene and/or gene transcript in a sample.

As used herein, a "nucleic acid molecule" is comprised of at least one base or one base pair, depending on whether it is single-stranded or double-stranded, respectively. Furthermore, a nucleic acid molecule may belong exclusively or chimerically to any group of nucleotide-containing molecules, as exemplified by, but not limited to, the following groups of nucleic acid molecules: RNA, DNA, genomic nucleic acids, non-genomic nucleic acids, naturally occurring and not naturally occurring nucleic acids, and synthetic nucleic acids. This includes, by way of non-limiting example, nucleic acids associated with any organelle, such as the mitochondria, ribosomal RNA, and nucleic acid molecules comprised chimerically of one or more components that are not naturally occurring along with naturally occurring components.

Additionally, a "nucleic acid molecule" may contain in part one or more non-nucleotide-based components as exemplified by, but not limited to, amino acids and sugars.

Thus, by way of example, but not limitation, a ribozyme that is in part nucleotide-based and in part protein-based is considered a "nucleic acid molecule".

In addition, by way of example, but not limitation, a nucleic acid molecule that is labeled with a detectable moiety, such as a radioactive or alternatively a non-radioactive label, is likewise considered a "nucleic acid molecule".

The terms "nucleic acid sequence coding for" or a "DNA coding sequence of" or a "nucleotide sequence encoding" a particular enzyme – as well as other synonymous terms – refer to a DNA sequence which is transcribed and translated into an enzyme when placed under the control of appropriate regulatory sequences. A "promotor sequence" is a DNA regulatory region capable of binding RNA polymerase in a cell and initiating transcription of a downstream (3' direction) coding sequence. The promoter is part of the DNA sequence. This sequence region has a start codon at its 3' terminus. The promoter sequence does include the minimum number of bases where elements necessary to initiate transcription at levels detectable above background. However, after the RNA polymerase binds the sequence and transcription is initiated at the start codon (3' terminus with a promoter), transcription proceeds downstream in the 3' direction. Within the promotor sequence will be found a transcription initiation site (conveniently defined by mapping with nuclease S1) as well as protein binding domains (consensus sequences) responsible for the binding of RNA polymerase.

The terms "nucleic acid encoding an enzyme (protein)" or "DNA encoding an enzyme (protein)" or "polynucleotide encoding an enzyme (protein)" and other synonymous terms encompasses a polynucleotide which includes only coding sequence for the enzyme as well as a polynucleotide which includes additional coding and/or non-coding sequence.

In one embodiment, a "specific nucleic acid molecule species" is defined by its chemical structure, as exemplified by, but not limited to, its primary sequence. In one exemplary embodiment, a specific "nucleic acid molecule species" is defined by a function of the nucleic acid species or by a function of a product derived from the nucleic acid species. Thus, by way of non-limiting example, a "specific nucleic acid molecule species" may be defined by one or more activities or properties attributable to it, including activities or properties attributable its expressed product.

The instant definition of "assembling a working nucleic acid sample into a nucleic acid library" includes the process of incorporating a nucleic acid sample into a vector-based

collection, such as by ligation into a vector and transformation of a host. A description of relevant vectors, hosts, and other reagents as well as specific non-limiting examples thereof are provided hereinafter. The instant definition of "assembling a working nucleic acid sample into a nucleic acid library" also includes the process of incorporating a nucleic acid sample into a non-vector-based collection, such as by ligation to adaptors. In one aspect, the adaptors can anneal to PCR primers to facilitate amplification by PCR.

Accordingly, in a non-limiting embodiment, a "nucleic acid library" is comprised of a vector-based collection of one or more nucleic acid molecules. In another embodiment a "nucleic acid library" is comprised of a non-vector-based collection of nucleic acid molecules. In yet another embodiment a "nucleic acid library" is comprised of a combined collection of nucleic acid molecules that is in part vector-based and in part non-vector-based. In one aspect, the collection of molecules comprising a library is searchable and separable according to individual nucleic acid molecule species.

The present invention provides a "nucleic acid construct" or alternatively a "nucleotide construct" or alternatively a "DNA construct". The term "construct" is used herein to describe a molecule, such as a polynucleotide (*e.g.*, a phytase polynucleotide) may optionally be chemically bonded to one or more additional molecular moieties, such as a vector, or parts of a vector. In a specific - but by no means limiting - aspect, a nucleotide construct is exemplified by a DNA expression DNA expression constructs suitable for the transformation of a host cell.

An "oligonucleotide" (or synonymously an "oligo") refers to either a single stranded polydeoxynucleotide or two complementary polydeoxynucleotide strands which may be chemically synthesized. Such synthetic oligonucleotides may or may not have a 5' phosphate. Those that do not will not ligate to another oligonucleotide without adding a phosphate with an ATP in the presence of a kinase. A synthetic oligonucleotide will ligate to a fragment that has not been dephosphorylated. To achieve polymerase-based amplification (such as with PCR), a "32-fold degenerate oligonucleotide that is comprised of, in series, at least a first homologous sequence, a degenerate N,N,G/T sequence, and a second homologous sequence" is mentioned. As used in this context, "homologous" is in reference to homology between the oligo and the parental polynucleotide that is subjected to the polymerase-based amplification.

A nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For instance, a promoter or enhancer is operably linked to a coding sequence if it increases the transcription of the coding sequence.

As used herein, the term "operably linked" refers to a linkage of polynucleotide elements in a functional relationship. A nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For instance, a promoter or enhancer is operably linked to a coding sequence if it affects the transcription of the coding sequence. Operably linked means that the DNA sequences being linked are typically contiguous and, where necessary to join two protein coding regions, contiguous and in reading frame. However, since enhancers generally function when separated from the promoter by several kilobases and intronic sequences may be of variable lengths, some polynucleotide elements may be operably linked but not contiguous.

A coding sequence is "operably linked to" another coding sequence when RNA polymerase will transcribe the two coding sequences into a single mRNA, which is then translated into a single polypeptide having amino acids derived from both coding sequences. The coding sequences need not be contiguous to one another so long as the expressed sequences are ultimately processed to produce the desired protein.

As used herein the term "parental polynucleotide set" is a set comprised of one or more distinct polynucleotide species. Usually this term fis used in reference to a progeny polynucleotide set which can be obtained by mutagenization of the parental set, in which case the terms "parental", "starting" and "template" are used interchangeably.

As used herein the term "physiological conditions" refers to temperature, pH, ionic strength, viscosity, and like biochemical parameters which are compatible with a viable organism, and/or which typically exist intracellularly in a viable cultured yeast cell or mammalian cell. For example, the intracellular conditions in a yeast cell grown under typical laboratory culture conditions are physiological conditions. Suitable *in vitro* reaction conditions for *in vitro* transcription cocktails are generally physiological conditions. In general, *in vitro* physiological conditions comprise 50-200 mM NaCl or KCl, pH 6.5-8.5, 20-45□C and 0.001-10 mM divalent cation (e.g., Mg^{++}, Ca^{++}); or about 150 mM NaCl or KCl, pH 7.2-7.6, 5 mM divalent cation, and often include 0.01-1.0 percent nonspecific protein (e.g., BSA). A non-ionic detergent (Tween, NP-40, Triton X-100) can often be present, usually at about 0.001 to 2%, typically 0.05-0.2% (v/v). Particular aqueous conditions may

be selected by the practitioner according to conventional methods. For general guidance, the following buffered aqueous conditions may be applicable: 10-250 mM NaCl, 5-50 mM Tris HCl, pH 5-8, with optional addition of divalent cation(s) and/or metal chelators and/or non-ionic detergents and/or membrane fractions and/or anti-foam agents and/or scintillants.

Standard convention (5' to 3') is used herein to describe the sequence of double standed polynucleotides.

The term "population" as used herein means a collection of components such as polynucleotides, portions or polynucleotides or proteins. A "mixed population: means a collection of components which belong to the same family of nucleic acids or proteins (i.e., are related) but which differ in their sequence (i.e., are not identical) and hence in their biological activity.

A molecule having a "pro-form" refers to a molecule that undergoes any combination of one or more covalent and noncovalent chemical modifications (e.g. glycosylation, proteolytic cleavage, dimerization or oligomerization, temperature-induced or pH-induced conformational change, association with a co-factor, etc.) en route to attain a more mature molecular form having a property difference (e.g. an increase in activity) in comparison with the reference pro-form molecule. When two or more chemical modification (e.g. two proteolytic cleavages, or a proteolytic cleavage and a deglycosylation) can be distinguished en route to the production of a mature molecule, the referemce precursor molecule may be termed a "pre-pro-form" molecule.

As used herein, the term "pseudorandom" refers to a set of sequences that have limited variability, such that, for example, the degree of residue variability at another position, but any pseudorandom position is allowed some degree of residue variation, however circumscribed.

The term "purified" denotes that a nucleic acid or protein gives rise to essentially one band in an electrophoretic gel. Particularly, it means that the nucleic acid or protein is at least about 50% pure, or at least about 85% pure, or at least about 99% pure.

"Quasi-repeated units", as used herein, refers to the repeats to be re-assorted and are by definition not identical. Indeed the method is proposed not only for practically identical encoding units produced by mutagenesis of the identical starting sequence, but also the reassortment of similar or related sequences which may diverge significantly in some

regions. Nevertheless, if the sequences contain sufficient homologies to be reassorted by this approach, they can be referred to as "quasi-repeated" units.

As used herein "random peptide library" refers to a set of polynucleotide sequences that encodes a set of random peptides, and to the set of random peptides encoded by those polynucleotide sequences, as well as the fusion proteins contain those random peptides.

As used herein, "random peptide sequence" refers to an amino acid sequence composed of two or more amino acid monomers and constructed by a stochastic or random process. A random peptide can include framework or scaffolding motifs, which may comprise invariant sequences.

As used herein, "receptor" refers to a molecule that has an affinity for a given ligand. Receptors can be naturally occurring or synthetic molecules. Receptors can be employed in an unaltered state or as aggregates with other species. Receptors can be attached, covalently or non-covalently, to a binding member, either directly or via a specific binding substance. Examples of receptors include, but are not limited to, antibodies, including monoclonal antibodies and antisera reactive with specific antigenic determinants (such as on viruses, cells, or other materials), cell membrane receptors, complex carbohydrates and glycoproteins, enzymes, and hormone receptors.

The term "recombinant" when used with reference to a cell indicates that the cell replicates a heterologous nucleic acid, or expresses a peptide or protein encoded by a heterologous nucleic acid. Recombinant cells can contain genes that are not found within the native (non-recombinant) form of the cell. Recombinant cells can also contain genes found in the native form of the cell wherein the genes are modified and re-introduced into the cell by artificial means. The term also encompasses cells that contain a nucleic acid endogenous to the cell that has been modified without removing the nucleic acid from the cell; such modifications include those obtained by gene replacement, site-specific mutation, and related techniques.

"Recombinant enzymes" refer to enzymes produced by recombinant DNA techniques, i.e., produced from cells transformed by an exogenous DNA construct encoding the desired enzyme. "Synthetic" enzymes are those prepared by chemical synthesis.

A "recombinant expression cassette" or simply an "expression cassette" is a nucleic acid construct, generated recombinantly or synthetically, with nucleic acid elements that are capable of effecting expression of a structural gené in hosts compatible with such sequences.

Expression cassettes include at least promoters and optionally, transcription termination signals. Typically, the recombinant expression cassette includes a nucleic acid to be transcribed (e.g., a nucleic acid encoding a desired polypeptide), and a promoter. Additional factors necessary or helpful in effecting expression may also be used as described herein. For example, an expression cassette can also include nucleotide sequences that encode a signal sequence that directs secretion of an expressed protein from the host cell. Transcription termination signals, enhancers, and other nucleic acid sequences that influence gene expression, can also be included in an expression cassette.

The term "related polynucleotides" means that regions or areas of the polynucleotides are identical and regions or areas of the polynucleotides are heterologous.

"Reductive reassortment", as used herein, refers to the increase in molecular diversity that is accrued through deletion (and/or insertion) events that are mediated by repeated sequences.

The following terms are used to describe the sequence relationships between two or more polynucleotides: "reference sequence," "comparison window," "sequence identity," "percentage of sequence identity," and "substantial identity."

A "reference sequence" is a defined sequence used as a basis for a sequence comparison; a reference sequence may be a subset of a larger sequence, for example, as a segment of a full-length cDNA or gene sequence given in a sequence listing, or may comprise a complete cDNA or gene sequence. Generally, a reference sequence is at least 20 nucleotides in length, frequently at least 25 nucleotides in length, and often at least 50 nucleotides in length. Since two polynucleotides may each (1) comprise a sequence (i.e., a portion of the complete polynucleotide sequence) that is similar between the two polynucleotides and (2) may further comprise a sequence that is divergent between the two polynucleotides, sequence comparisons between two (or more) polynucleotides are typically performed by comparing sequences of the two polynucleotides over a "comparison window" to identify and compare local regions of sequence similarity.

"Repetitive Index (RI)", as used herein, is the average number of copies of the quasi-repeated units contained in the cloning vector.

The term "restriction site" refers to a recognition sequence that is necessary for the manifestation of the action of a restriction enzyme, and includes a site of catalytic cleavage. It is appreciated that a site of cleavage may or may not be contained within a portion of a

restriction site that comprises a low ambiguity sequence (i.e. a sequence containing the principal determinant of the frequency of occurrence of the restriction site). Thus, in many cases, relevant restriction sites contain only a low ambiguity sequence with an internal cleavage site (e.g. G/AATTC in the EcoR I site) or an immediately adjacent cleavage site (e.g. /CCWGG in the EcoR II site). In other cases, relevant restriction enzymes [e.g. the Eco57 I site or CTGAAG(16/14)] contain a low ambiguity sequence (e.g. the CTGAAG sequence in the Eco57 I site) with an external cleavage site (e.g. in the N_{16} portion of the Eco57 I site). When an enzyme (e.g. a restriction enzyme) is said to "cleave" a polynucleotide, it is understood to mean that the restriction enzyme catalyzes or facilitates a cleavage of a polynucleotide.

The term "screening" describes, in general, a process that identifies optimal antigens. Several properties of the antigen can be used in selection and screening including antigen expression, folding, stability, immunogenicity and presence of epitopes from several related antigens. Selection is a form of screening in which identification and physical separation are achieved simultaneously by expression of a selection marker, which, in some genetic circumstances, allows cells expressing the marker to survive while other cells die (or vice versa). Screening markers include, for example, luciferase, beta-galactosidase and green fluorescent protein. Selection markers include drug and toxin resistance genes, and the like. Because of limitations in studying primary immune responses *in vitro*, *in vivo* studies are particularly useful screening methods. In these studies, the antigens are first introduced to test animals, and the immune responses are subsequently studied by analyzing protective immune responses or by studying the quality or strength of the induced immune response using lymphoid cells derived from the immunized animal. Although spontaneous selection can and does occur in the course of natural evolution, in the present methods selection is performed by man.

In a non-limiting aspect, a "selectable polynucleotide" is comprised of a 5' terminal region (or end region), an intermediate region (i.e. an internal or central region), and a 3' terminal region (or end region). As used in this aspect, a 5' terminal region is a region that is located towards a 5' polynucleotide terminus (or a 5' polynucleotide end); thus it is either partially or entirely in a 5' half of a polynucleotide. Likewise, a 3' terminal region is a region that is located towards a 3' polynucleotide terminus (or a 3' polynucleotide end); thus it is either partially or entirely in a 3' half of a polynucleotide. As used in this non-limiting

exemplification, there may be sequence overlap between any two regions or even among all three regions.

The term "sequence identity" means that two polynucleotide sequences are identical (i.e., on a nucleotide-by-nucleotide basis) over the window of comparison. The term "percentage of sequence identity" is calculated by comparing two optimally aligned sequences over the window of comparison, determining the number of positions at which the identical nucleic acid base (e.g., A, T, C, G, U, or I) occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison (i.e., the window size), and multiplying the result by 100 to yield the percentage of sequence identity. This "substantial identity", as used herein, denotes a characteristic of a polynucleotide sequence, wherein the polynucleotide comprises a sequence having at least 80 percent sequence identity, or at least 85 percent identity, often 90 to 95 percent sequence identity, and most commonly at least 99 percent sequence identity as compared to a reference sequence of a comparison window of at least 25-50 nucleotides, wherein the percentage of sequence identity is calculated by comparing the reference sequence to the polynucleotide sequence which may include deletions or additions which total 20 percent or less of the reference sequence over the window of comparison.

As known in the art "similarity" between two enzymes is determined by comparing the amino acid sequence and its conserved amino acid substitutes of one enzyme to the sequence of a second enzyme. Similarity may be determined by procedures which are well-known in the art, for example, a BLAST program (Basic Local Alignment Search Tool at the National Center for Biological Information).

As used herein, the term "single-chain antibody" refers to a polypeptide comprising a V_H domain and a V_L domain in polypeptide linkage, generally liked via a spacer peptide (e.g., $[Gly\text{-}Gly\text{-}Gly\text{-}Gly\text{-}Ser]_x$), and which may comprise additional amino acid sequences at the amino- and/or carboxy- termini. For example, a single-chain antibody may comprise a tether segment for linking to the encoding polynucleotide. As an example, a scFv is a single-chain antibody. Single-chain antibodies are generally proteins consisting of one or more polypeptide segments of at least 10 contiguous amino substantially encoded by genes of the immunoglobulin superfamily (e.g., see Williams and Barclay, 1989, pp. 361-368, which is incorporated herein by reference), most frequently encoded by a rodent, non-human primate,

avian, porcine bovine, ovine, goat, or human heavy chain or light chain gene sequence. A functional single-chain antibody generally contains a sufficient portion of an immunoglobulin superfamily gene product so as to retain the property of binding to a specific target molecule, typically a receptor or antigen (epitope).

The phrase "specifically (or selectively) binds to an antibody" or "specifically (or selectively) immunoreactive with", when referring to a protein or peptide, refers to a binding reaction which is determinative of the presence of the protein, or an epitope from the protein, in the presence of a heterogeneous population of proteins and other biologics. Thus, under designated immunoassay conditions, the specified antibodies bind to a particular protein and do not bind in a significant amount to other proteins present in the sample. The antibodies raised against a multivalent antigenic polypeptide will generally bind to the proteins from which one or more of the epitopes were obtained. Specific binding to an antibody under such conditions may require an antibody that is selected for its specificity for a particular protein. A variety of immunoassay formats may be used to select antibodies specifically immunoreactive with a particular protein. For example, solid-phase ELISA immunoassays, Western blots, or immunohistochemistry are routinely used to select monoclonal antibodies specifically immunoreactive with a protein. See Harlow and Lane (1988) Antibodies, A Laboratory Manual, Cold Spring Harbor Publications, New York "Harlow and Lane"), for a description of immunoassay formats and conditions that can be used to determine specific immunoreactivity. Typically a specific or selective reaction will be at least twice background signal or noise and more typically more than 10 to 100 times background.

The members of a pair of molecules (*e.g.*, an antibody-antigen pair or a nucleic acid pair) are said to "specifically bind" to each other if they bind to each other with greater affinity than to other, non-specific molecules. For example, an antibody raised against an antigen to which it binds more efficiently than to a non-specific protein can be described as specifically binding to the antigen. (Similarly, a nucleic acid probe can be described as specifically binding to a nucleic acid target if it forms a specific duplex with the target by base pairing interactions (see above).)

A "specific binding affinity" between two molecules, for example, a ligand and a receptor, means a preferential binding of one molecule for another in a mixture of molecules. The binding of the molecules can be considered specific if the binding affinity is about 1 X 10^4 M^{-1} to about 1 X $10^6 M^{-1}$ or greater.

"Specific hybridization" is defined herein as the formation of hybrids between a first polynucleotide and a second polynucleotide (e.g., a polynucleotide having a distinct but substantially identical sequence to the first polynucleotide), wherein substantially unrelated polynucleotide sequences do not form hybrids in the mixture.

The term "specific polynucleotide" means a polynucleotide having certain end points and having a certain nucleic acid sequence. Two polynucleotides wherein one polynucleotide has the identical sequence as a portion of the second polynucleotide but different ends comprises two different specific polynucleotides. The T_m is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Very stringent conditions are selected to be equal to the $T_{,,,}$ for a particular probe. An example of stringent hybridization conditions for hybridization of complementary nucleic acids which have more than 100 complementary residues on a filter in a Southern or northern blot is 50% formamide with I mg of heparin at 42'C, with the hybridization being carried out overnight.

"Stringent hybridization conditions" means hybridization will occur only if there is at least 90% identity, or at least 95% identity, or, at least 97% identity between the sequences. See, e.g., Sambrook et al, 1989. An example of highly "stringent" wash conditions is 0. 15M NaCl at 72'C for about 15 minutes. An example of stringent wash conditions is a 0.2x SSC wash at 65'C for 15 minutes (see, Sambrook, infra., for a description of SSC buffer). Often, a high stringency wash is preceded by a low stringency wash to remove background probe signal. An example medium stringency wash for a duplex of, e.g., more than 100 nucleotides, is 1x SSC at 45°C for 15 minutes. An example low stringency wash for a duplex of, e.g., more than 100 nucleotides, is 4-6x SSC at 40°C for 15 minutes. For short probes (e.g., about 10 to 50 nucleotides), stringent conditions typically involve salt concentrations of less than about 1.0 M Na^+ ion, typically about 0.01 to 1.0 M Na+ ion concentration (or other salts) at pH 7.0 to 8.3, and the temperature is typically at least about 30°C. Stringent conditions can also be achieved with the addition of destabilizing agents such as formamide. In general, a signal to noise ratio of 2x (or higher) than that observed for an unrelated probe in the particular hybridization assay indicates detection of a specific hybridization. Nucleic acids which do not hybridize to each other under stringent conditions are still substantially identical if the polypeptides which they encode T cell receptor polypeptides and major

histocompatibility molecules are substantially identical. This occurs, e.g., when a copy of a nucleic acid is created using the maximum codon degeneracy permitted by the genetic code.

"Stringent hybridization conditions" and "stringent hybridization wash conditions" in the context of nucleic acid hybridization experiments such as Southern and northern hybridizations are sequence dependent, and are different under different environmental parameters. Longer sequences hybridize specifically at higher temperatures. An extensive guide to the hybridization of nucleic acids is found in Tijssen (1993) Laboratory Techniques in Biochemistry and Molecular Biology-- Hybridization with Nucleic Acid Probes part I chapter 2 "Overview of principles of hybridization and the strategy of nucleic acid probe assays", Elsevier, New York. Generally, highly stringent hybridization and wash conditions are selected to be about 5°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. Typically, under "stringent conditions" a probe will hybridize to its target subsequence, but to no other sequences.

Also included in the invention are polypeptides having sequences that are "substantially identical" to the sequence of a phytase polypeptide, such as one of SEQ ID 1. A "substantially identical" amino acid sequence is a sequence that differs from a reference sequence only by conservative amino acid substitutions, for example, substitutions of one amino acid for another of the same class (*e.g.*, substitution of one hydrophobic amino acid, such as isoleucine, valine, leucine, or methionine, for another, or substitution of one polar amino acid for another, such as substitution of arginine for lysine, glutamic acid for aspartic acid, or glutamine for asparagine).

The phrase "substantially identical," in the context of two nucleic acids or polypeptides, refers to two or more sequences or subsequences that have at least 60%, or 80%, or 90-95% nucleotide or amino acid residue identity, when compared and aligned for maximum correspondence, as measured using one of the following sequence comparison algorithms or by visual inspection. In one aspect, the substantial identity exists over a region of the sequences that is at least about 50 residues in length or about 100 residues, or, the sequences are substantially identical over at least about 150 residues. In some embodiments, the sequences are substantially identical over the entire length of the coding regions.

A "subsequence" refers to a sequence of nucleic acids or amino acids that comprise a part of a longer sequence of nucleic acids or amino acids (e. g., polypeptide) respectively.

Additionally a "substantially identical" amino acid sequence is a sequence that differs from a reference sequence or by one or more non-conservative substitutions, deletions, or insertions, particularly when such a substitution occurs at a site that is not the active site the molecule, and provided that the polypeptide essentially retains its behavioural properties. For example, one or more amino acids can be deleted from a phytase polypeptide, resulting in modification of the structure of the polypeptide, without significantly altering its biological activity. For example, amino- or carboxyl-terminal amino acids that are not required for phytase biological activity can be removed. Such modifications can result in the development of smaller active phytase polypeptides.

The present invention provides a "substantially pure enzyme". The term "substantially pure enzyme" is used herein to describe a molecule, such as a polypeptide (*e.g.*, a phytase polypeptide, or a fragment thereof) that is substantially free of other proteins, lipids, carbohydrates, nucleic acids, and other biological materials with which it is naturally associated. For example, a substantially pure molecule, such as a polypeptide, can be at least 60%, by dry weight, the molecule of interest. The purity of the polypeptides can be determined using standard methods including, e.g., polyacrylamide gel electrophoresis (*e.g.*, SDS-PAGE), column chromatography (*e.g.*, high performance liquid chromatography (HPLC)), and amino-terminal amino acid sequence analysis.

As used herein, "substantially pure" means an object species is the predominant species present (i.e., on a molar basis it is more abundant than any other individual macromolecular species in the composition); alternatively, a substantially purified fraction is a composition wherein the object species comprises at least about 50 percent (on a molar basis) of all macromolecular species present. Generally, a substantially pure composition will comprise more than about 80 to 90 percent of all macromolecular species present in the composition. In one aspect, the object species is purified to essential homogeneity (contaminant species cannot be detected in the composition by conventional detection methods) wherein the composition consists essentially of a single macromolecular species. Solvent species, small molecules (<500 Daltons), and elemental ion species are not considered macromolecular species.

As used herein, the term "variable segment" refers to a portion of a nascent peptide which comprises a random, pseudorandom, or defined kernal sequence. A variable segment" refers to a portion of a nascent peptide which comprises a random pseudorandom, or defined

kernal sequence. A variable segment can comprise both variant and invariant residue positions, and the degree of residue variation at a variant residue position may be limited: both options are selected at the discretion of the practitioner. Typically, variable segments are about 5 to 20 amino acid residues in length (e.g., 8 to 10), although variable segments may be longer and may comprise antibody portions or receptor proteins, such as an antibody fragment, a nucleic acid binding protein, a receptor protein, and the like.

The term "wild-type" means that the polynucleotide does not comprise any mutations. A "wild type" protein means that the protein will be active at a level of activity found in nature and will comprise the amino acid sequence found in nature.

The term "working", as in "working sample", for example, is simply a sample with which one is working. Likewise, a "working molecule", for example is a molecule with which one is working.

Generating and Manipulating Nucleic Acids

The invention provides methods for generating variant antigen binding sites, antibodies and specific domains or fragments of antibodies (e.g., Fab or Fc domains) by manipulating a template nucleic acid, as described herein. The invention can be practiced in conjunction with any method or protocol or device known in the art, which are well described in the scientific and patent literature.

General Techniques

The nucleic acids used to practice this invention, whether RNA, cDNA, genomic DNA, vectors, viruses or hybrids thereof, may be isolated from a variety of sources, genetically engineered, amplified, and/or expressed/ generated recombinantly (recombinant polypeptides can be modified or immobilized to arrays in accordance with the invention). Any recombinant expression system can be used, including bacterial, mammalian, yeast, insect or plant cell expression systems.

Alternatively, these nucleic acids can be synthesized *in vitro* by well-known chemical synthesis techniques, as described in, e.g., Carruthers (1982) Cold Spring Harbor Symp. Quant. Biol. 47:411-418; Adams (1983) J. Am. Chem. Soc. 105:661; Belousov (1997) Nucleic Acids Res. 25:3440-3444; Frenkel (1995) Free Radic. Biol. Med. 19:373-380; Blommers (1994) Biochemistry 33:7886-7896; Narang (1979) Meth. Enzymol. 68:90; Brown (1979) Meth. Enzymol. 68:109; Beaucage (1981) Tetra. Lett. 22:1859; U.S. Patent No.

4,458,066. Double stranded DNA fragments may then be obtained either by synthesizing the complementary strand and annealing the strands together under appropriate conditions, or by adding the complementary strand using DNA polymerase with a primer sequence.

Techniques for the manipulation of nucleic acids, such as, e.g., subcloning, labeling probes (e.g., random-primer labeling using Klenow polymerase, nick translation, amplification), sequencing, hybridization and the like are well described in the scientific and patent literature, see, e.g., Sambrook, ed., MOLECULAR CLONING: A LABORATORY MANUAL (2ND ED.), Vols. 1-3, Cold Spring Harbor Laboratory, (1989); CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, Ausubel, ed. John Wiley & Sons, Inc., New York (1997); LABORATORY TECHNIQUES IN BIOCHEMISTRY AND MOLECULAR BIOLOGY: HYBRIDIZATION WITH NUCLEIC ACID PROBES, Part I. Theory and Nucleic Acid Preparation, Tijssen, ed. Elsevier, N.Y. (1993).

Another useful means of obtaining and manipulating nucleic acids used in the methods of the invention is to clone from genomic samples, and, if necessary, screen and re-clone inserts isolated (or amplified) from, e.g., genomic clones or cDNA clones or other sources of complete genomic DNA. Sources of genomic nucleic acid used in the methods and compositions of the invention include genomic or cDNA libraries contained in, or comprised entirely of, e.g., mammalian artificial chromosomes (see, e.g., Ascenzioni (1997) Cancer Lett. 118:135-142; U.S. Patent Nos. 5,721,118; 6,025,155) (including human artificial chromosomes, see, e.g., Warburton (1997) Nature 386:553-555; Roush (1997) Science 276:38-39; Rosenfeld (1997) Nat. Genet. 15:333-335); yeast artificial chromosomes (YAC); bacterial artificial chromosomes (BAC); P1 artificial chromosomes (see, e.g., Woon (1998) Genomics 50:306-316; Boren (1996) Genome Res. 6:1123-1130); PACs (a bacteriophage P1-derived vector, see, e.g., Ioannou (1994) Nature Genet. 6:84-89; Reid (1997) Genomics 43:366-375; Nothwang (1997) Genomics 41:370-378; Kern (1997) Biotechniques 23:120-124); cosmids, plasmids or cDNAs.

Amplification of Nucleic Acids

In one aspect of the invention, including methods using saturation mutagenesis, a template nucleic acid is amplified by an amplification reaction, such as a polymerase-based amplification, e.g., polymerase chain reaction (PCR). The amplification reaction is carried out using a 64-fold degenerate oligonucleotide for each codon to be mutagenized. The

skilled artisan can select and design suitable oligonucleotide amplification primers. Amplification methods are also well known in the art, and include, *e.g.*, polymerase chain reaction, PCR (see, e.g., PCR PROTOCOLS, A GUIDE TO METHODS AND APPLICATIONS, ed. Innis, Academic Press, N.Y. (1990) and PCR STRATEGIES (1995), ed. Innis, Academic Press, Inc., N.Y., ligase chain reaction (LCR) (see, e.g., Wu (1989) Genomics 4:560; Landegren (1988) Science 241:1077; Barringer (1990) Gene 89:117); transcription amplification (see, e.g., Kwoh (1989) Proc. Natl. Acad. Sci. USA 86:1173); and, self-sustained sequence replication (see, e.g., Guatelli (1990) Proc. Natl. Acad. Sci. USA 87:1874); Q Beta replicase amplification (see, e.g., Smith (1997) J. Clin. Microbiol. 35:1477-1491), automated Q-beta replicase amplification assay (see, e.g., Burg (1996) Mol. Cell. Probes 10:257-271) and other RNA polymerase mediated techniques (*e.g.*, NASBA, Cangene, Mississauga, Ontario); see also Berger (1987) Methods Enzymol. 152:307-316; Sambrook; Ausubel; U.S. Patent Nos. 4,683,195 and 4,683,202; Sooknanan (1995) Biotechnology 13:563-564.

Antibodies and Antigen Binding Sites

The invention provides methods for generating variant antigen binding sites, antibodies and specific domains or fragments of antibodies, e.g., Fab or Fc domains (defined above) by altering a template nucleic acid by saturation mutagenesis, an optimized directed evolution system, synthetic ligation reassembly, or a combination thereof. Antigen binding sites, antibodies or fragments thereof generated by these methods can be analyzed, e.g., screened for antigen binding activity (e.g., affinity, avidity) using a novel capillary array platform of the invention. All of an antibody sequence can be altered using one or more of these methods alone or in any order, or, subsequences or domains can be altered individually, and then can be reassembled in any order or orientation. For example, an Fc domain can be altered and screened for its ability to bind an Fc-cell surface receptor independently; the Fc segment can be religated to/ reassembled with an antigen binding domain afterwards.

The invention provides methods for generating variant nucleic acids from template sequences, such as antibody encoding sequences (e.g., genomic DNA or message) isolated from an organism, a cell or synthetically constructed. These nucleic acid sequences encoding for specific antigens, e.g., the template nucleic acids of the invention, can be generated by immunization followed by screening and isolation of the sequences encoding all or fragments

of antibodies that can specifically bind to that antigen. Methods of producing polyclonal and monoclonal antibodies are known to those of skill in the art and described in the scientific and patent literature, see, *e.g.*, Coligan, CURRENT PROTOCOLS IN IMMUNOLOGY, Wiley/Greene, NY (1991); Stites (eds.) BASIC AND CLINICAL IMMUNOLOGY (7th ed.) Lange Medical Publications, Los Altos, CA ("Stites"); Goding, MONOCLONAL ANTIBODIES: PRINCIPLES AND PRACTICE (2d ed.) Academic Press, New York, NY (1986); Kohler (1975) Nature 256:495; Harlow (1988) ANTIBODIES, A LABORATORY MANUAL, Cold Spring Harbor Publications, New York. Antibodies also can be generated *in vitro*, *e.g.*, using recombinant antibody binding site expressing phage display libraries, in addition to the traditional *in vivo* methods using animals. See, *e.g.*, Huse (1989) Science 246:1275; Ward (1989) Nature 341:544; Hoogenboom (1997) Trends Biotechnol. 15:62-70; Katz (1997) Annu. Rev. Biophys. Biomol. Struct. 26:27-45. Human antibodies can be generated in mice engineered to produce only human antibodies, as described by, e.g., U.S. Patent Nos. 5,877,397; 5,874,299; 5,789,650; and 5,939,598. B-cells from these mice can be immortalized using standard techniques (e.g., by fusing with an immortalizing cell line such as a myeloma or by manipulating such B-cells by other techniques to perpetuate a cell line) to produce a monoclonal human antibody-producing cell. See, e.g., U.S. Patent No. 5,916,771; 5,985,615.

For example, to generate human antibody encoding nucleic acids for a desired antigen, human lymphocytes can be inserted into an immunocompromised animal model, such as a SCID mouse. The animal is challenged with antigen one or more times and lymphocytes expressing an antibody specific for the antigen is isolated/ cloned. Alternatively, mice comprising human antibody genes that only express human antibodies can be used (discussed above).

Nucleic acid sequences (e.g., from cDNA libraries, isolated from human antibody producing mice, etc.) encoding desired antibodies can be cloned and further manipulated (e.g., to be used as templates in the methods of the invention). For example, if the antibody is of non-human origin, it can be "humanized" for eventual administration to patients. Methods for making chimeric, *e.g.*, "humanized," antibodies are well known in the art, see *e.g.*, U.S. Patent Nos. 5,811,522; 5,789,554; 5,861,155. Alternatively, recombinant antibodies can also be expressed by transient or stable expression vectors in mammalian, including human, cells and cell lines, as in Norderhaug (1997) J. Immunol. Methods 204:77-87; Boder (1997) Nat. Biotechnol. 15:553-557; see also U.S. Patent No. 5,976,833.

CHO cells lines that express "humanized" glycosylation patterns can be particularly useful, see, e.g., U.S. Patent No. 5,272,070.

The methods of the invention provide for "affinity enrichment" of an antibody or an antigen binding site. Antibody constant regions (e.g., Fc domains) can also be "affinity enriched" for their ability to specifically bind to an Fc receptor or a complement polypeptide. Very large sets, or libraries, of variant antibodies, including, e.g., CDRs, Fabs, Fcs, and single-chain antibodies, can be generated and screened for binding to ligand (e.g., antigen, complement, receptor, and the like). In one aspect, the variant polynucleotide is isolated and further manipulated by a method described herein, e.g., shuffled to recombine combinatorially the amino acid sequence of the selected polypeptides, peptide(s) or predetermined portions thereof. Thus, antibodies, antigen binding sites, Fc domains, and the like can be generated having a desired binding affinity for a molecule. The peptide or antibody can then be synthesized in bulk by conventional means for any suitable use (e.g., as a therapeutic pharmaceutical, a diagnostic agent, or as an *in vitro* reagent).

Saturation mutagenesis

This invention provides methods for generating variant antigen binding sites, antibodies and specific domains or fragments of antibodies (e.g., Fab or Fc domains), T cell receptor polypeptides and major histocompatibility molecules by altering template nucleic acids by saturation mutagenesis. In one aspect, codon primers containing a degenerate N,N,G/T sequence are used to introduce point mutations into a polynucleotide, so as to generate a set of progeny polypeptides in which a full range of single amino acid substitutions is represented at each amino acid position. These oligonucleotides can comprise a contiguous first homologous sequence, a degenerate N,N,G/T sequence, and, optionally, a second homologous sequence. The downstream progeny translational products from the use of such oligonucleotides include all possible amino acid changes at each amino acid site along the polypeptide, because the degeneracy of the N,N,G/T sequence includes codons for all 20 amino acids.

In one aspect, one such degenerate oligonucleotide (comprised of one degenerate N,N,G/T cassette) is used for subjecting each original codon in a parental polynucleotide template to a full range of codon substitutions. In another aspect, at least two degenerate N,N,G/T cassettes are used – either in the same oligonucleotide or not, for subjecting at least

two original codons in a parental polynucleotide template to a full range of codon substitutions. Thus, more than one N,N,G/T sequence can be contained in one oligonucleotide to introduce amino acid mutations at more than one site. This plurality of N,N,G/T sequences can be directly contiguous, or separated by one or more additional nucleotide sequence(s). In another aspect, oligonucleotides serviceable for introducing additions and deletions can be used either alone or in combination with the codons containing an N,N,G/T sequence, to introduce any combination or permutation of amino acid additions, deletions, and/or substitutions.

In one aspect, simultaneous mutagenesis of two or more contiguous amino acid positions is done using an oligonucleotide that contains contiguous N,N,G/T triplets, i.e. a degenerate (N,N,G/T)n sequence. In another aspect, degenerate cassettes having less degeneracy than the N,N,G/T sequence are used. For example, it may be desirable in some instances to use (e.g. in an oligo) a degenerate triplet sequence comprised of only one N, where said N can be in the first second or third position of the triplet. Any other bases including any combinations and permutations thereof can be used in the remaining two positions of the triplet. Alternatively, it may be desirable in some instances to use (e.g. in an oligo) a degenerate N,N,N triplet sequence.

In one aspect, use of degenerate N,N,G/T triplets allows for systematic and easy generation of a full range of possible natural amino acids (for a total of 20 amino acids) into each and every amino acid position in a polypeptide (in alternative aspects, the methods also include generation of less than all possible substitutions per amino acid residue, or codon, position). For example, for a 100 amino acid polypeptide, 2000 distinct species (i.e. 20 possible amino acids per position X 100 amino acid positions) can be generated. Through the use of an oligonucleotide or set of oligonucleotides containing a degenerate N,N,G/T triplet, 32 individual sequences can code for all 20 possible natural amino acids. Thus, in a reaction vessel in which a parental polynucleotide sequence is subjected to saturation mutagenesis using at least one such oligonucleotide, there are generated 32 distinct progeny polynucleotides encoding 20 distinct polypeptides. In contrast, the use of a non-degenerate oligonucleotide in site-directed mutagenesis leads to only one progeny polypeptide product per reaction vessel. Nondegenerate oligos can optionally be used in combination with degenerate primers disclosed; for example, nondegenerate oligonucleotides can be used to generate specific point mutations in a working polynucleotide. This provides one means to

generate specific silent point mutations, point mutations leading to corresponding amino acid changes, and point mutations that cause the generation of stop codons and the corresponding expression of polypeptide fragments.

In one aspect, each saturation mutagenesis reaction vessel contains polynucleotides encoding at least 20 progeny polypeptide molecules such that all 20 natural amino acids are represented at the one specific amino acid position corresponding to the codon position mutagenized in the parental polynucleotide (other aspects use less than all 20 natural combinations). The 32-fold degenerate progeny polypeptides generated from each saturation mutagenesis reaction vessel can be subjected to clonal amplification (e.g. cloned into a suitable host, e.g., *E. coli* host, using, e.g., an expression vector) and subjected to expression screening. When an individual progeny polypeptide is identified by screening to display a favorable change in property (when compared to the parental polypeptide, such as increased affinity or avidity to an antigen), it can be sequenced to identify the correspondingly favorable amino acid substitution contained therein.

In one aspect, upon mutagenizing each and every amino acid position in a parental polypeptide using saturation mutagenesis as disclosed herein, favorable amino acid changes may be identified at more than one amino acid position. One or more new progeny molecules can be generated that contain a combination of all or part of these favorable amino acid substitutions. For example, if 2 specific favorable amino acid changes are identified in each of 3 amino acid positions in a polypeptide, the permutations include 3 possibilities at each position (no change from the original amino acid, and each of two favorable changes) and 3 positions. Thus, there are 3 x 3 x 3 or 27 total possibilities, including 7 that were previously examined - 6 single point mutations (i.e. 2 at each of three positions) and no change at any position.

In yet another aspect, site-saturation mutagenesis can be used together with shuffling, chimerization, recombination and other mutagenizing processes, along with screening. This invention provides for the use of any mutagenizing process(es), including saturation mutagenesis, in an iterative manner. In one exemplification, the iterative use of any mutagenizing process(es) is used in combination with screening. Thus, in a non-limiting exemplification, this invention provides for the use of saturation mutagenesis in combination with additional mutagenization processes, such as process where two or more related

polynucleotides are introduced into a suitable host cell such that a hybrid polynucleotide is generated by recombination and reductive reassortment.

Optimized Directed Evolution System

This invention provides methods for generating variant antigen binding sites, antibodies and specific domains or fragments of antibodies (e.g., Fab or Fc domains), T cell receptor polypeptides and major histocompatibility molecules by manipulating a nucleic acid by an optimized directed evolution system. In one aspect, the invention further comprises mutagenizing a template nucleic acid, e.g., a nucleic acid encoding an antigen binding site, an antibody or fragment thereof, altered by saturation mutagenesis, by a method comprising an optimized directed evolution system. Optimized directed evolution is directed to the use of repeated cycles of reductive reassortment, recombination and selection that allow for the directed molecular evolution of nucleic acids through recombination. Optimized directed evolution allows generation of a large population of evolved chimeric sequences, wherein the generated population is significantly enriched for sequences that have a predetermined number of crossover events.

A crossover event is a point in a chimeric sequence where a shift in sequence occurs from one parental variant to another parental variant. Such a point is normally at the juncture of where oligonucleotides from two parents are ligated together to form a single sequence. This method allows calculation of the correct concentrations of oligonucleotide sequences so that the final chimeric population of sequences is enriched for the chosen number of crossover events. This provides more control over choosing chimeric variants having a predetermined number of crossover events.

In addition, this method provides a convenient means for exploring a tremendous amount of the possible protein variant space in comparison to other systems. Previously, if one generated, for example, 10^{13} chimeric molecules during a reaction, it would be extremely difficult to test such a high number of chimeric variants for a particular activity. Moreover, a significant portion of the progeny population would have a very high number of crossover events which resulted in proteins that were less likely to have increased levels of a particular activity. By using these methods, the population of chimerics molecules can be enriched for those variants that have a particular number of crossover events. Thus, although one can still generate 10^{13} chimeric molecules during a reaction, each of the molecules chosen for further

analysis most likely has, for example, only three crossover events. Because the resulting progeny population can be skewed to have a predetermined number of crossover events, the boundaries on the functional variety between the chimeric molecules is reduced. This provides a more manageable number of variables when calculating which oligonucleotide from the original parental polynucleotides might be responsible for affecting a particular trait, such as antigen binding.

One method for creating a chimeric progeny polynucleotide sequence is to create oligonucleotides corresponding to fragments or portions of each parental sequence. Each oligonucleotide can include a unique region of overlap so that mixing the oligonucleotides together results in a new variant that has each oligonucleotide fragment assembled in the correct order. Additional information can also be found in U.S. Patent Application Number 09/332,835 entitled "Synthetic Ligation Reassembly in Directed Evolution" and filed on June 14, 1999, the disclosure of which has been incorporated by reference in its entirety. The number of oligonucleotides generated for each parental variant bears a relationship to the total number of resulting crossovers in the chimeric molecule that is ultimately created. For example, three parental nucleotide sequence variants might be provided to undergo a ligation reaction in order to find a chimeric variant having, for example, greater activity at high temperature. As one example, a set of 50 oligonucleotide sequences can be generated corresponding to each portions of each parental variant. Accordingly, during the ligation reassembly process there could be up to 50 crossover events within each of the chimeric sequences. The probability that each of the generated chimeric polynucleotides will contain oligonucleotides from each parental variant in alternating order is very low. If each oligonucleotide fragment is present in the ligation reaction in the same molar quantity it is likely that in some positions oligonucleotides from the same parental polynucleotide will ligate next to one another and thus not result in a crossover event. If the concentration of each oligonucleotide from each parent is kept constant during any ligation step in this example, there is a 1/3 chance (assuming 3 parents) that an oligonucleotide from the same parental variant will ligate within the chimeric sequence and produce no crossover.

Accordingly, a probability density function (PDF) can be determined to predict the population of crossover events that are likely to occur during each step in a ligation reaction given a set number of parental variants, a number of oligonucleotides corresponding to each variant, and the concentrations of each variant during each step in the ligation reaction. The

statistics and mathematics behind determining the PDF is described below. By utilizing these methods, one can calculate such a probability density function, and thus enrich the chimeric progeny population for a predetermined number of crossover events resulting from a particular ligation reaction. Moreover, a target number of crossover events can be predetermined, and the system then programmed to calculate the starting quantities of each parental oligonucleotide during each step in the ligation reaction to result in a probability density function that centers on the predetermined number of crossover events.

These methods are directed to the use of repeated cycles of reductive reassortment, recombination and selection that allow for the directed molecular evolution of a nucleic acid encoding an antigen binding site through recombination. This system allows generation of a large population of evolved chimeric sequences, wherein the generated population is significantly enriched for sequences that have a predetermined number of crossover events. A crossover event is a point in a chimeric sequence where a shift in sequence occurs from one parental variant to another parental variant. Such a point is normally at the juncture of where oligonucleotides from two parents are ligated together to form a single sequence. The method allows calculation of the correct concentrations of oligonucleotide sequences so that the final chimeric population of sequences is enriched for the chosen number of crossover events. This provides more control over choosing chimeric variants having a predetermined number of crossover events.

In addition, these methods provide a convenient means for exploring a tremendous amount of the possible protein variant space in comparison to other systems. By using the methods described herein, the population of chimerics molecules can be enriched for those variants that have a particular number of crossover events. Thus, although one can still generate 10^{13} chimeric molecules during a reaction, each of the molecules chosen for further analysis most likely has, for example, only three crossover events. Because the resulting progeny population can be skewed to have a predetermined number of crossover events, the boundaries on the functional variety between the chimeric molecules is reduced. This provides a more manageable number of variables when calculating which oligonucleotide from the original parental polynucleotides might be responsible for affecting a particular trait.

In one aspect, the method creates a chimeric progeny polynucleotide sequence by creating oligonucleotides corresponding to fragments or portions of each parental sequence (e.g., first antigen binding site, or template, sequence). Each oligonucleotide can include a

unique region of overlap so that mixing the oligonucleotides together results in a new variant that has each oligonucleotide fragment assembled in the correct order. Additional information can also be found in U.S. Patent Application No. 09/332,835 entitled "Synthetic Ligation Reassembly in Directed Evolution" and filed on June 14, 1999.

The number of oligonucleotides generated for each parental variant bears a relationship to the total number of resulting crossovers in the chimeric molecule that is ultimately created. For example, three parental nucleotide sequence variants might be provided to undergo a ligation reaction in order to find a chimeric variant having, for example, greater activity at high temperature. As one example, a set of 50 oligonucleotide sequences can be generated corresponding to each portions of each parental variant. Accordingly, during the ligation reassembly process there could be up to 50 crossover events within each of the chimeric sequences. The probability that each of the generated chimeric polynucleotides will contain oligonucleotides from each parental variant in alternating order is very low. If each oligonucleotide fragment is present in the ligation reaction in the same molar quantity it is likely that in some positions oligonucleotides from the same parental polynucleotide will ligate next to one another and thus not result in a crossover event. If the concentration of each oligonucleotide from each parent is kept constant during any ligation step in this example, there is a 1/3 chance (assuming 3 parents) that a oligonucleotide from the same parental variant will ligate within the chimeric sequence and produce no crossover.

Accordingly, a probability density function (PDF) can be determined to predict the population of crossover events that are likely to occur during each step in a ligation reaction given a set number of parental variants, a number of oligonucleotides corresponding to each variant, and the concentrations of each variant during each step in the ligation reaction. The statistics and mathematics behind determining the PDF is described below. One can calculate such a probability density function, and thus enrich the chimeric progeny population for a predetermined number of crossover events resulting from a particular ligation reaction. Moreover, a target number of crossover events can be predetermined, and the system then programmed to calculate the starting quantities of each parental oligonucleotide during each step in the ligation reaction to result in a probability density function that centers on the predetermined number of crossover events.

Determining Crossover Events

Embodiments of the invention include a system and software that receive a desired crossover probability density function (PDF), the number of parent genes to be reassembled, and the number of fragments in the reassembly as inputs. The output of this program is a "fragment PDF" that can be used to determine a recipe for producing reassembled genes, and the estimated crossover PDF of those genes. The processing described herein can be performed in MATLAB® (The Mathworks, Natick, Massachusetts) a programming language and development environment for technical computing.

Computer System

One aspect of the system is the computer system that carries out the methods described herein. In one aspect, the computer system is a conventional personal computer such as those based on an Intel microprocessor and running a Windows operating system. The output of the computer system is a fragment PDF that can be used as a recipe for producing reassembled progeny genes, and the estimated crossover PDF of those genes. The processing described herein can be performed by a personal computer using the MATLAB® programming language and development environment. The invention is not limited to any particular hardware or software configuration. For example, computers based on other well-known microprocessors and running operating system software such as UNIX, Linux, MacOS and others are contemplated.

Iterative Reassembly

In various aspects, the methods generate sets of chimeric nucleic acid and protein molecules and then screen those molecules for a particular activity, such as the ability to bind to a desired antigen. The invention is not limited to only a single round of screening. For example, a second round of screening can take place if nucleotide sequencing indicates that all of the chimeric progeny antibody polynucleotides having an increased affinity or specificity have a particular parental oligonucleotide in common. Based on this determination, a second round of reassembly can take place that enriches for progeny having that oligonucleotide. This can be done by, for example, not adding the corresponding oligonucleotide sequences from the other parental polynucleotides into the ligation reassembly reactions. Thus, the only oligonucleotide that can be ligated into each gene will be the desired oligonucleotide.

Similarly, if it is determined that a particular oligonucleotide has no affect at all on the desired trait (e.g., affinity for antigen), it can be removed as a variable by synthesizing larger parental oligonucleotides that include the sequence to be removed. Since incorporating the sequence within a larger sequence prevents any crossover events, there will no longer be any variation of this sequence in the progeny polynucleotides. This iterative practice of determining which oligonucleotides are most related to the desired trait, and which are unrelated, allows more efficient exploration all of the possible protein variants that might be provide a particular trait or activity.

Automated Control of Reactions

The process of generating any of the reactions of the methods of the invention can be automated with the assistance of robotic instruments. For example, a TECAN GENESIS™ programmable robot made by Tecan Corporation (Hombrechtikon, Switzerland) can be interfaced with a computer that determines the quantities of each oligonucleotide fragment to yield a resulting PDF. By linking a computer system that determines the proper quantities of each oligonucleotide to an automated robot, a complete ligation reassembly system is produced. Data links through serial or other interfaces will allow the data files generated from the ligation reassembly calculations to be forwarded in the proper format for the robotic system to automatically begin allocating the proper quantities of each oligonucleotide fragment into a reaction tube.

Thus, one aspect of the invention is an automated system for generating nucleic acid sequences that encode variant antigen binding sites, such as variant antibodies having increased affinity to desired antigen. The automated system includes a plurality of oligonucleotide fragments derived from a series of nucleic acid sequence variants, wherein said fragments are configured to join one another at unique overhangs. The system also has a data input field configured to store a target number of crossover events in for each of the variant sequences. Within the system is also a prediction module configured to determine the quantity of each of the fragments to admix together so that mixing the fragments results in a population of progeny molecules that are enriched for crossover events corresponding to the target number. The system also provides a robotic arm linked to the prediction module through a communication interface for automatically mixing the fragments in the determined quantities.

Mutagenized Oligonucleotides

While the optimized directed evolution method can use oligonucleotides that have a 100% fidelity to their parent polynucleotide sequence, this level of fidelity is not required. For example, if a set of three related parental polynucleotides are chosen to undergo ligation reassembly in order to create, e.g., an antibody having increase affinity to a desired antigen, a set of oligonucleotides having unique overlapping regions can be synthesized by conventional methods. However a set of mutagenized oligonucleotides could also be synthesized. These mutagenized oligonucleotides can be designed to encode silent, conservative, or non-conservative amino acids.

The choice to enter a silent mutation might be made to, for example, add a region of nucleotide homology two fragments, but not affect the final translated protein. A non-conservative or conservative substitution is made to determine how such a change alters the function of the resultant polypeptide. This can be done if, for example, it is determined that mutations in one particular oligonucleotide fragment were responsible for increasing the activity of a peptide. By synthesizing mutagenized oligonucleotides (e.g.: those having a different nucleotide sequence than their parent), one can explore, in a controlled manner, how resulting modifications to the peptide sequence affect the activity of the peptide, e.g., affinity to a desired antigen.

Another method for creating variants of a nucleic acid sequence using mutagenized fragments includes first aligning a plurality of nucleic acid sequences to determine demarcation sites within the variants that are conserved in a majority of said variants, but not conserved in all of said variants. A set of first sequence fragments of the conserved nucleic acid sequences are then generated, wherein the fragments bind to one another at the demarcation sites. A second set of fragments of the not conserved nucleic acid sequences are then generated by, for example, a nucleic acid synthesizer. However, the not conserved, sequences are generated to have mutations at their demarcation site so that the second fragments have the same nucleotide sequence at the demarcation sites as said first fragments. This allows the not conserved sequences to still hybridize during the ligation reaction to the other parental sequences. Once the fragments are generated, a desired number of crossover events can be selected for each of the variants. The quantity of each of the first and second fragments is then calculated so that a ligation/incubation reaction between the calculated

quantities of the first and second fragments will result in progeny molecules having the desired number of crossover events.

Synthetic Ligation Reassembly (SLR)

This invention provides methods for generating variant antigen binding sites, antibodies and specific domains or fragments of antibodies (e.g., Fab or Fc domains) by altering template nucleic acids by synthetic ligation reassembly. SLR is a method of ligating oligonucleotide fragments together non-stochastically. This method differs from stochastic oligonucleotide shuffling in that the nucleic acid building blocks are not shuffled, concatenated or chimerized randomly, but rather are assembled non-stochastically. The SLRs used in the methods of the invention do not depend on the presence of high levels of homology between polynucleotides to be rearranged. Thus, this method can be used to non-stochastically generate libraries (or sets) of progeny molecules comprised of over 10^{100} different chimeras. SLR can be used to generate libraries comprised of over 10^{1000} different progeny chimeras. Thus, aspects of the present invention include non-stochastic methods of producing a set of finalized chimeric nucleic acid molecules (e.g., nucleic acids encoding antibodies or fragments thereof) having an overall assembly order that is chosen by design. This method includes the steps of generating by design a plurality of specific nucleic acid building blocks having serviceable mutually compatible ligatable ends, and assembling these nucleic acid building blocks, such that a designed overall assembly order is achieved.

The mutually compatible ligatable ends of the nucleic acid building blocks to be assembled are considered to be "serviceable" for this type of ordered assembly if they enable the building blocks to be coupled in predetermined orders. Thus the overall assembly order in which the nucleic acid building blocks can be coupled is specified by the design of the ligatable ends. If more than one assembly step is to be used, then the overall assembly order in which the nucleic acid building blocks can be coupled is also specified by the sequential order of the assembly step(s). In one aspect, the annealed building pieces are treated with an enzyme, such as a ligase (e.g. T4 DNA ligase), to achieve covalent bonding of the building pieces.

In one aspect, the design of the oligonucleotide building blocks is obtained by analyzing a set of progenitor nucleic acid sequence templates (parents, such as antibody coding sequences) that serve as a basis for producing a progeny set of finalized chimeric

polynucleotide molecules (e.g., variant antibodies). These parental oligonucleotide templates thus serve as a source of sequence information that aids in the design of the nucleic acid building blocks that are to be mutagenized, e.g., chimerized or shuffled.

In one aspect, a chimerization of a set, or family, of related genes and their encoded set, or family, of polypeptides is provided. The encoded products can be antibodies or fragments or subsequences thereof, such as Fc or Fab domains, antigen binding sites, CDRs, and the like.

In one aspect of this method, the sequences of a plurality of parental nucleic acid templates are aligned in order to select one or more demarcation points. The demarcation points can be located at an area of homology, and are comprised of one or more nucleotides. These demarcation points can be shared by at least two of the progenitor templates. The demarcation points can thereby be used to delineate the boundaries of oligonucleotide building blocks to be generated in order to rearrange the parental polynucleotides. The demarcation points identified and selected in the progenitor molecules serve as potential chimerization points in the assembly of the final chimeric progeny molecules. A demarcation point can be an area of homology (comprised of at least one homologous nucleotide base) shared by at least two parental polynucleotide sequences. Alternatively, a demarcation point can be an area of homology that is shared by at least half of the parental polynucleotide sequences, or, it can be an area of homology that is shared by at least two thirds of the parental polynucleotide sequences. In one aspect, a serviceable demarcation points is an area of homology that is shared by at least three fourths of the parental polynucleotide sequences, or, it can be shared by at almost all of the parental polynucleotide sequences. In one aspect, a demarcation point is an area of homology that is shared by all of the parental polynucleotide sequences.

In one aspect, a ligation reassembly process is performed exhaustively in order to generate an exhaustive library of progeny chimeric polynucleotides. In other words, all possible ordered combinations of the nucleic acid building blocks are represented in the set of finalized chimeric nucleic acid molecules. At the same time, in another embodiment, the assembly order (i.e. the order of assembly of each building block in the 5' to 3 sequence of each finalized chimeric nucleic acid) in each combination is by design (or non-stochastic) as described above. Because of the non-stochastic nature of this invention, the possibility of unwanted side products is greatly reduced.

In another aspect, the ligation reassembly method is performed systematically. For example, the method is performed in order to generate a systematically compartmentalized library of progeny molecules, with compartments that can be screened systematically, e.g. one by one. In other words this invention provides that, through the selective and judicious use of specific nucleic acid building blocks, coupled with the selective and judicious use of sequentially stepped assembly reactions, a design can be achieved where specific sets of progeny products are made in each of several reaction vessels. This allows a systematic examination and screening procedure to be performed. Thus, these methods allow a potentially very large number of progeny molecules to be examined systematically in smaller groups.

Because of its ability to perform chimerizations in a manner that is highly flexible yet exhaustive and systematic as well, particularly when there is a low level of homology among the progenitor molecules, these methods provide for the generation of a library (or set) comprised of a large number of progeny molecules. Because of the non-stochastic nature of the instant ligation reassembly invention, the progeny molecules generated can comprise a library of finalized chimeric nucleic acid molecules having an overall assembly order that is chosen by design. In alternative aspects, sets, or a library, of generated progeny molecules (nucleic acids or polypeptides) comprises greater than 10^3 different progeny molecular species, greater than 10^5 different progeny molecular species, greater than 10^{10} different progeny molecular species, greater than 10^{15} different progeny molecular species, greater than 10^{20} different progeny molecular species, greater than 10^{30} different progeny molecular species, greater than 10^{40} different progeny molecular species, greater than 10^{50} different progeny molecular species, greater than 10^{60} different progeny molecular species, greater than 10^{70} different progeny molecular species, greater than 10^{80} different progeny molecular species, or greater than 10^{100} different progeny molecular species, greater than 10^{110} different progeny molecular species, greater than 10^{120} different progeny molecular species, greater than 10^{130} different progeny molecular species, greater than 10^{140} different progeny molecular species, greater than 10^{150} different progeny molecular species, greater than 10^{175} different progeny molecular species, greater than 10^{200} different progeny molecular species, greater than 10^{300} different progeny molecular species, greater than 10^{400} different progeny molecular species, greater than 10^{500} different progeny molecular species, and greater than 10^{1000} different progeny molecular species.

The saturation mutagenesis and optimized directed evolution methods also can be used to generate these amounts of different progeny molecular species.

In one aspect, a set of finalized chimeric nucleic acid molecules, produced as described herein, comprises a polynucleotide encoding a polypeptide. According to another aspect, this polynucleotide is a gene, which may be a man-made gene. According to another aspect, this polynucleotide is an antibody or a fragment thereof.

It is appreciated that the invention provides freedom of choice and control regarding the selection of demarcation points, the size and number of the nucleic acid building blocks, and the size and design of the couplings. It is appreciated, furthermore, that the requirement for intermolecular homology is highly relaxed for the operability of this invention. In fact, demarcation points can even be chosen in areas of little or no intermolecular homology. For example, because of codon wobble, i.e. the degeneracy of codons, nucleotide substitutions can be introduced into nucleic acid building blocks without altering the amino acid originally encoded in the corresponding progenitor template. Alternatively, a codon can be altered such that the coding for an originally amino acid is altered. This invention provides that such substitutions can be introduced into the nucleic acid building block in order to increase the incidence of intermolecularly homologous demarcation points and thus to allow an increased number of couplings to be achieved among the building blocks, which in turn allows a greater number of progeny chimeric molecules to be generated.

In another aspect, the synthetic nature of the step in which the building blocks are generated allows the design and introduction of nucleotides (e.g., one or more nucleotides, which may be, for example, codons or introns or regulatory sequences) that can later be optionally removed in an *in vitro* process (e.g. by mutageneis) or in an *in vivo* process (e.g. by utilizing the gene splicing ability of a host organism). It is appreciated that in many instances the introduction of these nucleotides may also be desirable for many other reasons in addition to the potential benefit of creating a serviceable demarcation point.

Thus, according to another aspect, a nucleic acid building block can be used to introduce an intron. Thus, functional introns may be introduced into a man-made gene manufactured according to the methods described herein. In addition, functional introns may be introduced into a man-made antibody gene of this invention. Accordingly, these methods provide for the generation of a chimeric polynucleotide that is a man-made gene containing one (or more) artificially introduced intron(s). The artificially introduced intron(s) are

functional in one or more host cells for gene splicing much in the way that naturally-occurring introns serve functionally in gene splicing. A process of producing man-made intron-containing polynucleotides to be introduced into host organisms for recombination and/or splicing is also contemplated.

Screening methodologies

In alternative aspects of the methods of the invention, the set of progeny nucleic acids, e.g., antibody-, Fc-, antigen binding site- encoding polynucleotides, T cell receptor polypeptides and major histocompatibility molecules are expressed. These polypeptides can be expressed to screen for their ability to bind a ligand, e.g., an antigen (if, for example, affinity maturation of an antibody is desired), or, a receptor or a complement molecule (e.g., for Fc domains). Any method of expression or screening known in the art can be used.

The displayed peptide or polypeptide sequences can be of varying lengths, e.g., from 3-5000 amino acids long or longer, from 5-100 amino acids long, or from about 8-15 amino acids long. A set, or library, can comprise library members having varying lengths of displayed peptide sequence, or may comprise library members having a fixed length of displayed peptide sequence. Exemplary display methods include methods for *in vitro* and *in vivo* display of single-chain antibodies, such as nascent scFv on polysomes or scfv displayed on phage, which enable large-scale screening of scfv libraries having broad diversity of variable region sequences and binding specificities.

The present invention also provides random, pseudorandom, and defined sequence framework nucleic acid and polypeptide libraries and methods for generating and screening those libraries to identify useful compounds (e.g., antibodies, including single-chain antibodies, Fc, and the like) that bind to receptor molecules or antigens or epitopes of interest. The random, pseudorandom, and defined sequence framework peptides can be produced from libraries of peptide library members that comprise displayed peptides or displayed single-chain antibodies attached to a polynucleotide template from which the displayed peptide was synthesized. The mode of attachment may vary according to the specific embodiment of the invention selected, and can include encapsulation in a phage particle or incorporation in a cell.

Screening with Capillary Arrays

In one aspect of the invention, the variant nucleic acids are expressed and the generated polypeptides, e.g., antibodies, including antigen binding sites, CDRs, or Fab or Fc, are screened for their ability to specifically bind a molecule, e.g., an antigen, a complement molecule, an Fc receptor, by a method comprising a capillary array, such as GIGAMATRIX™, Diversa Corporation, San Diego, CA.

The capillary arrays of the invention provide a system and method for holding and screening samples. In one aspect of the capillary array invention, a sample screening apparatus includes a plurality of capillaries formed into an array of adjacent capillaries, wherein each capillary comprises at least one wall defining a lumen for retaining a sample. The apparatus further includes interstitial material disposed between adjacent capillaries in the array, and one or more reference indicia formed within of the interstitial material. According to another aspect of the invention, a capillary for screening a sample, wherein the capillary is adapted for being bound in an array of capillaries, includes a first wall defining a lumen for retaining the sample, and a second wall formed of a filtering material, for filtering excitation energy provided to the lumen to excite the sample.

In one aspect, the invention provides a method for incubating a biomolecule of interest (e.g., the antibody or fragment thereof, or, a ligand, such as an epitope or antigen, to be screened) includes the steps of introducing a first component into at least a portion of a capillary of a capillary array, wherein each capillary of the capillary array comprises at least one wall defining a lumen for retaining the first component, and introducing an air bubble into the capillary behind the first component. The method further includes the step of introducing a second component into the capillary, wherein the second component is separated from the first component by the air bubble. In another aspect, a sample of interest is introduced as a first liquid labeled with a detectable particle into a capillary of a capillary array, wherein each capillary of the capillary array comprises at least one wall defining a lumen for retaining the first liquid and the detectable particle, and wherein the at least one wall is coated with a binding material for binding the detectable particle to the at least one wall. The method can further include removing the first liquid from the capillary tube, wherein the bound detectable particle is maintained within the capillary, and introducing a second liquid into the capillary tube.

In one aspect, variant polypeptide, e.g., the antibody or fragment thereof, is immobilized onto the capillary array (or other device if another screening method is used)

(i.e., the antibody is in "solid phase"). Alternatively, the ligand, such as an epitope or antigen, to be screened, immobilized onto the device, e.g., the capillary array (i.e., the ligand, such as an antigen, is in "solid phase").

In one aspect, the capillary array includes a plurality of individual capillaries comprising at least one outer wall defining a lumen. The outer wall of the capillary can be one or more walls fused together. Similarly, the wall can define a lumen that is cylindrical, square, hexagonal or any other geometric shape so long as the walls form a lumen for retention of a liquid or sample. The capillaries of the capillary array can be held together in close proximity to form a planar structure. The capillaries can be bound together, by being fused (*e.g.*, where the capillaries are made of glass), glued, bonded, or clamped side-by-side. The capillary array can be formed of any number of individual capillaries, for example, a range from 100 to 4,000,000 capillaries. A capillary array can form a microtiter plate having about 100,000 or more individual capillaries bound together.

The capillaries can be formed with an aspect ratio of 50:1. In one aspect, each capillary has a length of approximately 10mm, and an internal diameter of the lumen of approximately 200μm. However, other aspect ratios are possible, and range from 10:1 to well over 1000:1. Accordingly, individual capillaries have an inner diameter that ranges from 10-500μm. A capillary having an internal diameter of 200 μm and a length of 1 cm has a volume of about 0.3 μl. The length and width of each capillary can be based on a desired volume and other characteristics, such as evaporation rate, etc. The capillary array can have a density of 500 to more than 1,000 capillaries per cm^2, or about 5 capillaries per mm^2. The capillary array can be formed to a width or diameter of about 0.5-20 mm and a height or thickness of 0.05 to about 10 cm. The capillary array can have a thickness of about 0.1 to about 5 cm.

In one aspect of the methods of the invention, one or more particles (comprising antigen/ ligand or antibody, depending on which is to be in solid phase for the screening) are introduced into each capillary for screening. Suitable particles include cells, cell clones, and other biological matter, chemical beads, or any other particulate matter. The capillaries containing particles of interest can be exposed with various types of substances for screening for an activity of interest, e.g., antibody binding to antigen, Fc binding to complement, and the like. A chemical solution containing new particles can be introduced to cause a combining event with other chemical beads already introduced into one or more capillaries. The particles and resulting activity of interest are screened and analyzed using the capillary array. In one aspect,

the activity produces optical energy within the capillary, which can act as a waveguide for guiding the light energy to an analyzer.

The capillaries can be made according to various manufacturing techniques. In one aspect, the capillaries are manufactured using a hollow-drawn technique. A cylindrical, or other hollow shape, portion of glass is drawn out to continually longer lengths according to known techniques. The glass is drawn to a desired diameter and then cut into portions of a specific length to form a capillary according to the invention. Then, a number of individual capillaries are bound together in an array. In an alternative aspect, a glass etching process is used. A solid tube of glass can be drawn out to a particular width, and cut into portions of a specific length. Then, each solid tube portion is center-etched with acid to form a capillary. The tubes can be bound before or after the etch process. A large number of materials can be suitably used to form a capillary array according to the invention and depending on the manufacturing technique used, including without limitation, glass, metal, semiconductors such as silicon, quartz, ceramics, or various polymers and plastics including, among others, polyethylene, polystyrene, and polypropylene. The internal walls of the capillary array, or portions thereof, may be coated or silanized to modify their surface properties. For example, the hydrophilicity or hydrophobicity may be altered to promote or reduce wicking or capillary action, respectively. The coating material includes, for example, ligands such as avidin, streptavidin, antibodies, antigens, and other molecules having specific binding affinity or which can withstand thermal or chemical sterilization.

A capillary array may optionally include reference indicia for providing a positional or alignment reference. The reference indicia may be formed of a pad of glass extending from the surface of the capillary array, or embedded in the interstitial material. In one aspect, the reference indicia are provided at one or more corners of a microtiter plate formed by the capillary array. A corner of the plate or set of capillaries may be removed, and replaced with the reference indicia. The reference indicia may also be formed at spaced intervals along a capillary array, to provide an indication of a subset of capillaries.

The capillary can include a first wall defining a lumen and a second wall surrounding the first wall. In one aspect, the second wall has a lower index of refraction than the first wall. In one aspect, the first wall is a sleeve glass having a high index of refraction, forming a waveguide in which light from excited fluorophores travels. The second wall can be black EMA glass, having a low index of refraction, forming a cladding around the first wall against

which light is refracted and directed along the first wall for total internal reflection within the capillary. The second wall can thus be made with any material that reduces the "cross-talk" or diffusion of light between adjacent capillaries. Alternatively, the inside surface of the first wall can be coated with a reflective substance to form a mirror, or mirror-like structure, for specular reflection within the lumen. Many different materials can be used in forming the first and second walls, creating different indices of refraction for desired purposes. A filtering material can be formed around the lumen to filter energy to and from the lumen. In one aspect, the inner wall of the first wall of each capillary of the array, or portion of the array, is coated with the filtering material. In another aspect, the second wall includes the filtering material. For instance, the second wall can be formed of the filtering material, such as filter glass for example, or in one aspect, the second wall is EMA glass that is doped with an appropriate amount of filtering material. The filtering material can be formed of a color other than black and tuned for a desired excitation/ emission filtering characteristic. The filtering material can allow transmission of excitation energy into the lumen, and blocks emission energy from the lumen except through one or more openings at either end of the capillary. Excitation energy is illustrated as a solid line, while emission energy is indicated by a broken line. When the second wall is formed with a filtering material, certain wavelengths of light representing excitation energy are allowed through to the lumen, and other wavelengths of light representing emission energy are blocked from exiting, except as directed within and along the first wall. The entire capillary array, or a portion thereof, can be tuned to a specific individual wavelength or group of wavelengths, for filtering different bands of light in an excitation and detection process.

In one aspect, during use, an excitation light is directed into the lumen contacting a particle (discussed above) and exciting a reporter fluorescent material causing emission of light. The emitted light travels the length of the capillary until it reaches a detector. If the second wall is black EMA glass, emitted light cannot cross contaminate adjacent capillary tubes in a capillary array. In addition, the black EMA glass refracts and directs the emitted light towards either end of the capillary tube thus increasing the signal detected by an optical detector (*e.g.*, a CCD camera and the like).

In a detection process using a capillary array of the invention, an optical detection system is aligned with the array, which is then scanned for one or more bright spots, representing either a fluorescence or luminescence associated with a "positive." The term

"positive" refers to the presence of an activity of interest. Again, the activity can be a chemical event, or a biological event. In one aspect, a capillary array is immersed or contacted with a container containing particles or molecules of interest. The particles can be cells, clones, molecules or compounds (e.g., antibodies or fragments thereof, antigens, and the like) suspended in a liquid. The liquid is wicked into the capillary tubes by capillary action. The natural wicking that occurs as a result of capillary forces obviates the need for pumping equipment and liquid dispensers. A substrate for measuring biological activity (*e.g.*, antibody affinity) can be contacted with the particles either before or after introduction of the particles into the capillaries in the capillary array. The substrate can include clones of a cell of interest, for example. The substrate can be introduced simultaneously into the capillaries by placing an open end of the capillaries in the container containing a mixture of the particle-bearing liquid and the substrate. Alternatively, the particle-bearing liquid may be wicked a portion of the way into the capillaries, and then the substrate is wicked into a remaining portion of the capillaries.

The mixture in the capillaries can then be incubated for producing a desired activity, e.g., a binding event, such as antibody binding to antigen, Fc to complement, and the like. The incubation can be for a specific period of time and at an appropriate temperature or to allow the substrate to permeabilize the cell membrane to produce an optically detectable signal, or for a period of time and at a temperature for optimum binding activity. The incubation can be performed, for example, by placing the capillary array in a humidified incubator or at ambient temperature in an apparatus containing a water source to ensure reduced evaporation within the capillary tubes. The evaporative flow rate may be reduced by increasing the humidity (*e.g.*, by placing the capillary array in a humidified chamber). The evaporation rate can also be reduced by capping the capillaries with an oil, wax, membrane or the like. Alternatively, a high molecular weight fluid such as various alcohols, or molecules capable of forming a molecular monolayer, bilayers or other thin films (*e.g.*, fatty acids), or various oils (*e.g.*, mineral oil) can be used to reduce evaporation.

In one aspect, a first fluid is wicked into the capillary according to methods described above. The capillary containing the substrate solution is then introduced to a fluid bath containing a second liquid. The second liquid may or may not be the same as the first. For instance, the first liquid may contain particles from which an activity is screened. The particles are suspended in liquid within the lumen, and gradually migrate toward the top of

the lumen in the direction of the flow of liquid through the capillary. The width of the lumen at the open end of the capillary can be sized to provide a particular surface area of liquid at the top of the lumen, for controlling the amount and rate of evaporation of the liquid mixture. By controlling the environment near the fluid bath, the first liquid from within the capillary will evaporate, and will be replenished by the second liquid from the fluid bath. The amount of evaporation is balanced against possible diffusion of the contents of the capillary into the liquid, and against possible mechanical mixing of the capillary contents with the liquid due to vibration and pressure changes. The greater the length of the capillary, the less the capillary contents will mix with the liquid and be subject to diffusion. The greater the width of the lumen, the larger the amount of mechanical mixing. Therefore, the temperature and humidity level in the surrounding environment may be adjusted to produce the desired evaporative cycle, and the lumen width is sized to minimize mechanical mixing, in addition to produce a desired evaporation rate. The non-submersed open end of the capillary may also be capped to create a vacuum force for holding the capillary contents within the capillary, and minimizing mechanical mixing and diffusion of the contents within the liquid. However when capped, the capillary will not experience evaporation. A relatively high humidity level of the environment will slow the rate of evaporation and keep more liquid within the capillary. If a heat differential between the environment and capillary array exceeds a certain level, however, evaporating or other liquid can condense on a top surface of tightly-packed capillaries of the capillary array. The outer edge surface of the capillary walls can be a planar surface. The wall of the capillary can be glass, the outer edge surface of the capillary wall can be polished glass.

In order to minimize condensation, a hydrophobic coating can be provided over the outer edge surface of the capillary walls. The coating can reduce the tendency for water or other liquid to accumulate near the outer edge surface of the capillary wall. In one aspect, the hydrophobic coating is TEFLON™. In one configuration, the coating covers only the outer edge surfaces of the capillary walls. In another configuration, the coating can be formed over both the interstitial material and the outer edge surfaces of the capillary walls. Another advantage of a hydrophobic coating over the outer edge surface of the capillary tubes is during the initial wicking process, some fluidic material in the form of droplets will tend to stick to the surface in which the fluid is introduced. Therefore, the coating minimizes extraneous fluid from

forming on the surface of a capillary array, dispensing with a need to shake or knock the extraneous fluid from the surface.

In some aspects, it is a goal to achieve a certain concentration of particles of interest, e.g., antigens, antibodies. A process of dilution, may be used to achieve a particular concentration, or series of dilutions, of particles. In one aspect employing dilution, a bolus of a first component is wicked into a capillary by capillary action until only a portion of the capillary is filled. In one aspect, pressure is applied at one end of the capillary to prevent the first component from wicking into the entire capillary. The end of the capillary may be completely or partially capped to provide the pressure. An amount of air is then introduced into the capillary adjacent the first component. The air can be introduced by any number of processes. One such process includes moving the first component in one direction within the capillary until a suitable amount of the air (84) is introduced behind the first component. Further movement of the first component by a pulling and/or pushing pressure causes a piston-like action by the first component on the air. The capillary or capillary array is then contacted to a second component. The second component can be pulled into the capillary by the piston-like action created by movement of the first component until a suitable amount of the second component is provided in the capillary, separated from the first component by the air. One of the first or second components may contain one or more particles of interest, and the other of the components may be a developer of the particles for causing an activity of interest. The capillary or capillary array can then be incubated for a period of time to allow the first and second components to reach an optimal temperature, or for a sufficient time to allow cell growth for example. The air-bubble separating the two components can be disrupted in order to allow mix the two components together and initialize the desired activity. In one aspect, pressure is applied to either one of the components or to the entire capillary to collapse the bubble.

One of the components may contain paramagnetic beads or particles. The paramagnetic beads can be used to disrupt the air bubble and/or mix the contents of the capillary tube or capillary array. For example, paramagnetic beads can be magnetically attracted from one location in each to another location. The paramagnetic beads are attracted by magnetic fields formed in proximity to the capillary or capillary array. By alternating or adjusting the location of the magnetic field with respect to each capillary, the paramagnetic beads will move within each capillary to mix the liquid within the capillary in which the beads are suspended. Mixing

the liquid can improve cell growth by increasing aeration of the cells. This aspect also improves consistency and detectability of the liquid sample among the capillaries.

In another aspect, a method of forming a multi-component assay includes providing one or more capsules of a second component within a first component. The second component capsules can have an outer layer of a substance that melts or dissolves at a predetermined temperature, thereby releasing the second component into the first component and combining particles among the components. One such substance is a thermally activated enzyme. Alternatively, a "release on command" mechanism that is configured to release the second component upon a predetermined event or condition may also be used.

In another aspect, recombinant clones containing a reporter construct or a substrate are wicked into the capillary tubes of the capillary array. In this aspect, it is not necessary to add a substrate as the reporter construct or substrate contained in the clone can be readily detected using techniques known in the art. For example, a clone containing a reporter construct such as green fluorescent protein can be detected by exposing the clone or substrate within the clone to a wavelength of light that induces fluorescence. Such reporter constructs can be implemented to respond to various conditions or upon exposure to various physical stimuli (including light and heat). In addition, various compounds can be screened in a sample using similar techniques. For example, an antibody or antigen detectably labeled with a florescent molecule can be readily detected within a capillary tube of a capillary array.

In yet another aspect, instead of dilution, a fluorescence-activated cell sorter (FACS) is used to separate and isolate particles or clones for delivery into the capillary array; thus, one or more clones per capillary tube can be precisely achieved.

Some assays may require an exchange of media within the capillary. In a media exchange process, a first liquid containing the particles is wicked into a capillary. The first liquid is removed, and replaced with a second liquid while the particles remain suspended within the capillary. Addition of the second liquid to the capillary and contact with the particles can initialize an activity, such as an assay, for example. The media exchange process may include a mechanism by which the particles in the capillary are physically maintained in the capillary while the first liquid is removed. In one aspect, the inner walls of the capillary array are coated with antibodies to which an antigen, e.g., a cell, can bind. Then, the first liquid is removed, while the antigen remains bound to the antibodies, and the second liquid is wicked into the capillary. The second liquid could be adapted to cause the antigens to unbind if

desirable. In an alternative aspect, one or more walls of the capillary can be magnetized. The particles are also magnetized and attracted to the walls. In still another aspect, magnetized particles are attracted and held against one side of the capillary upon application of a magnetic field near that side.

The capillary array can be analyzed for identification of capillaries having a detectable signal, such as an optical signal (*e.g.*, fluorescence), by a detector capable of detecting a change in light production or light transmission, for example. Detection may be performed using an illumination source that provides fluorescence excitation to each of the capillaries in the array, and a photodetector that detects resulting emission from the fluorescence excitation. Suitable illumination sources include, without limitation, a laser, incandescent bulb, light emitting diode (LED), and arch discharge. Suitable photodetectors include, without limitation, a photodiode array, a charge-coupled device (CCD), or charge injection device (CID). In one aspect, a detection system includes a laser source that produces a laser beam. The laser beam can be directed into a beam expander configured to produce a wider or less divergent beam for exciting the array of capillaries. Suitable laser sources include argon or ion lasers. A cooled CCD can be used.

If light is generated by, for example, enzymatic activation of a fluorescent substrate, it can be detected by an appropriate light detector or detectors positioned adjacent to the apparatus of the invention. The light detector may be, for example, film, a photomultiplier tube, photodiode, avalanche photo diode, CCD or other light detector or camera. The light detector may be a single detector to detect sequential emissions, such as a scanning laser. Or, the light detector may include a plurality of separate detectors to detect and spatially resolve simultaneous emissions at single or multiple wavelengths of emitted light. The light emitted and detected may be visible light or may be emitted as non-visible radiation such as infrared or ultraviolet radiation. A thermal detector may be used to detect an infrared emission. The detector or detectors may be stationary or movable. The emitted light or other radiation, such as illumination, may be channeled to the detector or detectors by means of lenses, mirrors and fiber optic light guides or light conduits (single, multiple, fixed, or moveable) positioned on or adjacent to at least one surface of the capillary array.

The photodetector can comprise a CCD, CID or an array of photodiode elements. Detection of a position of one or more capillaries having an optical signal can then be determined from the optical input from each element. Alternatively, the array may be scanned

by a scanning confocal or phase-contrast fluorescence microscope or the like, where the array is, for example, carried on a movable stage for movement in a X-Y plane as the capillaries in the array are successively aligned with the beam to determine the capillary array positions at which an optical signal is detected. A CCD camera or the like can be used in conjunction with the microscope. The detection system can be computer-automated for rapid screening and recovery. A telecentric lens can be used for detection. Magnification of the telecentric lens is adjusted to match the camera to the plane of view of the capillary array.

Where a chromogenic substrate is used, the change in the absorbance spectrum can be measured, such as by using a spectrophotometer or the like. Such measurements are usually difficult when dealing with a low-volume liquid because the optical path length is short. However, the capillary approach of the present invention permits small volumes of liquid to have long optical path lengths (*e.g.*, longitudinally along the capillary tube), thereby providing the ability to measure absorbance changes using conventional techniques.

In another aspects, binding or other activity is detected by using various electromagnetic detection devices, including, for example, optical, magnetic and thermal detection. In yet another aspects, radioactivity can be detected within a capillary tube using detection methods known in the art. The radiation can be detected at either end of the capillary tube. Other detection modes include, without limitation, luminescence, fluorescence polarization, time-resolved fluorescence. Luminescence detection includes detecting emitted light that is produced by a chemical or physiological process associated with a sample molecule or cell. Fluorescence polarization detection includes excitation of the contents of the lumen with polarized light. Under such environment, a fluorophore emits polarized light for a particular molecule. However, the emitting molecule can be moving and changing its angle of orientation, and the polarized light emission could become random.

Time-resolved fluorescence includes reading the fluorescence at a predetermined time after excitation. For a long-life fluorophore, the molecule is flashed with excitation energy, which produces emissions from the fluorophore as well as from other particles within the substrate. Emissions from the other particles causes background fluorescence. The background fluorescence normally has a short lifetime relative to the long-life emission from the fluorophore. The emission can be read after excitation is complete, at a time when all background fluorescence usually has short lifetime, and during a time in which the long-life

fluorophores continues to fluoresce. Time-resolved fluorescence can be a technique for suppressing background fluorescent activity.

A fluid within a capillary will usually form a meniscus at each end. Any light entering the capillary will be deflected toward the wall, except for paraxial rays, which enter the meniscus curvature at its center. The paraxial rays create a small bright spot in middle of capillary, representing the small amount of light that makes it through. Measurement of the bright spot provides an opportunity to measure how much light is being absorbed on its way through. In one aspect, a detection system includes the use of two different wavelengths. A ratio between a first and a second wavelength indicates how much light is absorbed in the capillary. Alternatively, two images of the capillary can be taken, and a difference between them can be used to ascertain a differential absorbance of a chemical within the capillary. In absorbance detection, only light in the center of the lumen can travels through the capillary. However, if at least one meniscus flattened, the optical efficiency is improved. The meniscus can be kept flat under a number of circumstances, such as in the evaporative wick cycle. The fluid bath can be contained in a clear, light-passing container, and the light source can be directed through the fluid bath into the capillary.

Recovery of putative hits (e.g., antigen binding to antibody) producing a detectable or optical signal can be facilitated by using position feedback from the detection system to automate positioning of a recovery device (*e.g.*, a needle pipette tip or capillary tube). In this example, a needle is selected and connected to recovery mechanism. A support table supports a microtiter plate capillary array and a light source. The light source is used with a camera assembly to find a location in the Z-axis of a needle connected to the recovery mechanism. The support table moves in the axis of X and Y, to place the capillary array underneath the needle, where the capillary array contains a "hit." The recovery mechanism then guides the needle to a capillary containing a "hit" by overlapping the tip of the needle with the capillary containing the "hit," in the Z direction, until the tip of the needle engages the capillary opening. In order to avoid damage to the capillary itself the needle may be attached to a spring or be of a material that flexes. Once in contact with the opening of the capillary the sample can be aspirated or expelled from the capillary.

In an exemplary recovery technique, a single camera is used for determining a location of a recovery tool, such as the tip of a needle, in the Z-plane. The Z-plane determination can be accomplished using an auto-focus algorithm, or proximity sensor used in conjunction with the

camera. Once the proximity of the recovery tool in Z is known, an image processing function can be executed to determine a precise location of the recovery tool in X and Y. In one aspect, the recovery tool is back-lit to aid the image processing. Once the X and Y coordinate locations are known, the capillary array can be moved in X and Y relative to the precise location of the recovery tool, which can be moved along the Z axis for coupling with a target capillary. In an alternative aspect of a recovery technique, two or more cameras are used for determining a location of the recovery tool. For instance, a first camera can determine X and Z coordinate locations of the recovery tool, such as the X, Z location of a needle tip. A second camera can determine Y and Z coordinate locations of the recovery tool. The two sets of coordinates can then be multiplexed for a complete X,Y,Z coordinate location. Next, the movement of the capillary array relative to the recovery tool can be executed.

The sample can be expelled by, for example, injecting a blast of inert gas into the capillary and collecting the ejected sample in a collection device at the opposite end of the capillary. The diameter of the collection device can be larger than or equal to the diameter of the capillary. The collected sample can then be further processed by, for example, extracting polynucleotides, proteins or by growing the clone in culture. In another aspect, the sample is aspirated by use of a vacuum. In this aspect, the needle contacts, or nearly contacts, the capillary opening and the sample is "vacuumed" or aspirated from the capillary tube onto or into a collection device. The collection device may be a microfuge tube or a filter located proximal to the opening of the needle. Suitable collection devices include a microfuge tube, a capillary tube, microtiter plate, cell culture plate, and the like. The delivery of the sample can be accomplished by forcing another media, air or other fluid through the filter in the reverse direction. The sample can also be expelled from a capillary by a sample ejector. In one aspect, the ejector is a jet system where sample fluid at one end of the capillary tube is subjected to a high temperature, causing fluid at the other end of the capillary tube to eject out. The heating of fluid can be accomplished mechanically, by applying a heated probe directly into one end of a capillary tube. The heated probe can seal the one end, heats fluid in contact with the probe, and expels fluid out the other end of the capillary tube . The heating and expulsion may also be accomplished electronically. For instance, in an embodiment of the jet system, at least one wall of a capillary tube is metalized. A heating element is placed in direct contact with one end of the wall. The heating element may completely close off the one end, or partially close the one end. The heating element charges up the metalized wall, which generates heat within

the fluid. The heating element can be an electricity source, such as a voltage source, or a current source. In still yet another embodiment of a jet system, a laser applies heat pulses to the fluid at one end of the capillary tube. Other systems for expelling fluid from a capillary tube of the invention can be used. An electric field may be created in or near the fluid to create an electrophoretic reaction, which causes the fluid to move according to electromotive force created by the electric field. An electric field may also assist in guiding a heated probe or electrically charged element to a target location near the fluid. An electromagnetic field may also be used. In one aspect, the capillary tube contains, in addition to the fluid, magnetically charged particles to help move the fluid out of the capillary array.

GENERAL CONSIDERATIONS & FORMATS FOR RECOMBINATION

Component modules provides genetic vaccine with the acquisition of or improvement in a useful property or characteristic.

The present invention provides multicomponent genetic vaccines that include one or more component modules, each of which provides the genetic vaccine with the acquisition of or an improvement in a property or characteristic useful in genetic vaccination.

The invention provides significant advantages over previously used genetic vaccines. Through use of a multicomponent vaccine, one can obtain an immune response that is particularly effective for a particular application. A multicomponent genetic vaccine can, for example, contain a component that is optimized for optimal antigen expression, as well as a component that confers improved activation of cytotoxic T lymphocytes (CTLs) by enhancing the presentation of the antigen on dendritic cell MHC Class I molecules. Additional examples are described herein.

The invention provides a new approach to vaccine development, which is termed "antigen library immunization." No other technologies are available for generating libraries of related antigens or optimizing known protective antigens. The most powerful previously existing methods for identification of vaccine antigens, such as high throughput sequencing or expression library immunization, only explore the sequence space provided by the pathogen genome. These approaches are likely to be insufficient, because they generally only target single pathogen strains, and because natural evolution has directed pathogens to downregulate their own immunogenicity. In contrast, the immunization protocols of the invention, which use experimentally evolved (e.g. by polynucleotide reassembly &/or

polynucleotide site-saturation mutagenesis) antigen libraries, provide a means to identify novel antigen sequences. Those antigens that are most protective can be selected from these pools by *in vivo* challenge models. Antigen library immunization dramatically expands the diversity of available immunogen sequences, and therefore, these antigen chimera libraries can also provide means to defend against newly emerging pathogen variants of the future. The methods of the invention enable the identification of individual chimeric antigens that provide efficient protection against a variety of existing pathogens, providing improved vaccines for troops and civilian populations.

The methods of the invention provide an evolution-based approach, such as stochastic (e.g. polynucleotide shuffling & interrupted synthesis) and non-stochastic polynucleotide reassembly in particular, that is an optimal approach to improve the immunogenicity of many types of antigens. For example, the methods provide means of obtaining optimized cancer antigens useful for preventing and treating malignant diseases. Furthermore, an increasing number of self-antigens, causing autoimmune diseases, and allergens, causing atopy, allergy and asthma, have been characterized. The immunogenicity and manufacturing of these antigens can likewise be improved with the methods of this invention.

The antigen library immunization methods of the invention provide a means by which one can obtain a recombinant antigen that has improved ability to induce an immune response to a pathogenic agent. A "pathogenic agent" refers to an organism or virus that is capable of infecting a host cell. Pathogenic agents typically include and/or encode a molecule, usually a polypeptide, that is immunogenic in that an immune response is raised against the immunogenic polypeptide. Often, the immune response raised against an immunogenic polypeptide from one serotype of the pathogenic agent is not capable of recognizing, and thus protecting against, a different serotype of the pathogenic agent, or other related pathogenic agents. In other situations, the polypeptide produced by a pathogenic agent is not produced in sufficient amounts, or is not sufficiently immunogenic, for the infected host to raise an effective immune response against the pathogenic agent.

These problems are overcome by the methods of the invention, which typically involve reassembling (&/or subjecting to one or more directed evolution methods described herein) two or more forms of a nucleic acid that encode a polypeptide of the pathogenic agent, or antigen involved in another disease or condition. These reassembly methods, including stochastic (e.g. polynucleotide shuffling & interrupted synthesis) and non-

stochastic polynucleotide reassembly, use as substrates forms of the nucleic acid that differ from each other in two or more nucleotides, so a library of recombinant nucleic acids results. The library is then screened to identify at least one optimized recombinant nucleic acid that encodes an optimized recombinant antigen that has improved ability to induce an immune response to the pathogenic agent or other condition.

The resulting recombinant antigens often are chimeric in that they are recognized by antibodies (Abs) reacting against multiple pathogen strains, and generally can also elicit broad spectrum immune responses. Specific neutralizing antibodies are known to mediate protection against several pathogens of interest, although additional mechanisms, such as cytotoxic T lymphocytes, are likely to be involved. The concept of chimeric, multivalent antigens inducing broadly reacting antibody responses is further illustrated herein.

In alternative embodiments, the different forms of the nucleic acids that encode antigenic polypeptides are obtained from members of a family of related pathogenic agents.

This scheme of performing stochastic (e.g. polynucleotide shuffling & interrupted synthesis) and non-stochastic polynucleotide reassembly using nucleic acids from different organisms is shown schematically herein. Therefore, these stochastic (e.g. polynucleotide shuffling & interrupted synthesis) and non-stochastic polynucleotide reassembly methods provide an effective approach to generate multivalent, crossprotective antigens. The methods are useful for obtaining individual chimeras that effectively protect against most or all pathogen variants.

Moreover, immunizations using entire libraries or pools of experimentally evolved (e.g. by polynucleotide reassembly &/or polynucleotide site-saturation mutagenesis) antigen chimeras can also result in identification of chimeric antigens that protect against pathogen variants that were not included in the starting population of antigens (for example, protection against strain C by experimentally evolved (e.g. by polynucleotide reassembly &/or polynucleotide site-saturation mutagenesis) library of chimeras/mutants of strains A and B).

Accordingly, the antigen library immunization approach enables the development of immunogenic polypeptides that can induce immune responses against poorly characterized, newly emerging pathogen variants.

Sequence reassembly (&/or one or more additional directed evolution methods described herein) can be achieved in many different formats and permutations of formats, as described in further detail below. These formats share some common principles. For

example, the targets for modification vary in different applications, as does the property sought to be acquired or improved. Examples of candidate targets for acquisition of a property or improvement in a property include genes that encode proteins which have immunogenic and/or toxigenic activity when introduced into a host organism.

The methods use at least two variant forms of a starting target. The variant forms of candidate substrates can show substantial sequence or secondary structural similarity with each other, but they should also differ in at least one, or, alternatively, in at least two positions. The initial diversity between forms can be the result of natural variation, e.g., the different variant forms (homologs) are obtained from different individuals or strains of an organism, or constitute related sequences from the same organism (e.g. , allelic variations), or constitute homologs from different organisms (interspecific variants).

Alternatively, initial diversity can be induced, e.g., the variant forms can be generated by error-prone transcription, such as an error-prone PCR or use of a polymerase which lacks proof-reading activity (see, Liao (1990) Gene 88:107-111), of the first variant form, or, by replication of the first form in a mutator strain (mutator host cells are discussed in further detail below, and are generally well known). A mutator strain can include any mutants in any organism impaired in the functions of mismatch repair. These include mutant gene products of mutS, mutT, mutH, mutL, ovrD, dcm, vsr, umuC, umuD, sbcB, recJ, etc. The impairment is achieved by genetic mutation, allelic replacement, selective inhibition by an added reagent such as a small compound or an expressed antisense RNA, or other techniques. Impairment can be of the genes noted, or of homologous genes in any organism. Other methods of generating initial diversity include methods well known to those of skill in the art, including, for example, treatment of a nucleic acid with a chemical or other mutagen, through spontaneous mutation, and by inducing an error-prone repair system (e.g., SOS) in a cell that contains the nucleic acid. The initial diversity between substrates is greatly augmented in subsequent steps of reassembly (&/or one or more additional directed evolution methods described herein) for library generation.

Properties involved in immunogenicity

Polynucleotide sequences that can positively or negatively affect the immunogenicity of an antigen encoded by the polynucleotide are often scattered throughout the entire antigen gene. Several of these factors are shown diagrammatically herein. By reassembling (&/or subjecting to one or more directed evolution methods described herein) different forms of

polynucleotide that encode the antigen using stochastic (e.g. polynucleotide shuffling & interrupted synthesis) and non-stochastic polynucleotide reassembly, followed by selection for those chimeric polynucleotides that encode an antigen that can induce an improved immune response, one can obtain primarily sequences that have a positive influence on antigen immunogenicity. Those sequences that negatively affect antigen immunogenicity are eliminated. One need not know the particular sequences involved.

The present invention provides methods for obtaining polynucleotide sequences that, either directly or indirectly (i.e., through encoding a polypeptide), can modulate an immune response when present on a genetic vaccine vector. In another embodiment, the invention provides methods for optimizing the transport and presentation of antigens. The optimized immunomodulatory polynucleotides obtained using the methods of the invention are particularly suited for use in conjunction with vaccines, including genetic vaccines. One of the advantages of genetic vaccines is that one can incorporate genes encoding immunomodulatory molecules, such as cytokines, costimulatory molecules, and molecules that improve antigen transport and presentation into the genetic vaccine vectors. This provides opportunities to modulate immune responses that are induced against the antigens expressed by the genetic vaccines.

Obtaining components for use in genetic vaccines that are more effective through the creation of a library, the screening of the library, and the use of recombinant nucleic acids that exhibit improved properties.

In additional embodiments, the present invention provides methods of obtaining components for use in genetic vaccines, including the multicomponent vaccines, that are more effective in conferring a desired immune response property upon a genetic vaccine. The methods involve creating a library of recombinant nucleic acids and screening the library to identify those library members that exhibits an enhanced capacity to confer a desired property upon a genetic vaccine. Those recombinant nucleic acids that exhibit improved properties can be used as components in a genetic vaccine, either directly as a polynucleotide or as a protein that is obtained by expression of the component nucleic acid.

Improvement goals

The properties or characteristics that are acquired or improved by the methods of the invention vary widely, and, of course depend on the choice of substrate. For antibodies, they include "affinity maturation," or, the generation of antibodies with a higher affinity for an

antigen. For T cell receptors, this can include an increased or decreased affinity for antigen, as presented by a major histocompatibility complex molecule. For genetic vaccines, improvement goals include higher titer, more stable expression, improved stability, higher specificity targeting, higher or lower frequency of integration, reduced immunogenicity of the vector or an expression product thereof, increased immunogenicity of the antigen, higher expression of gene products, and the like. Other properties for which optimization is desired include the tailoring of an immune response to be most effective for a particular application. Examples of genetic vaccine components are shown, described &/or referenced herein (including incorporated by reference). Two or more components can be included in a single vector molecule, or each component can be present in a genetic vaccine formulation as a separate molecule.

Sequence reassembly (&/or one or more additional directed evolution methods described herein) can be achieved through different formats which share some common principles

In the methods of the invention, at least two variant forms of a nucleic acid are reassembled (&/or subjected to one or more directed evolution methods described herein) to produce a library of recombinant nucleic acids, which is then screened to identify at least one recombinant component that is optimized for the particular vaccine property. Often, improvements are achieved after one round of reassembly (&/or one or more additional directed evolution methods described herein) and selection. Sequence reassembly (&/or one or more additional directed evolution methods described herein) can be achieved in many different formats and permutations of formats, as described in further detail below. These formats share some common principles. A family of nucleic acid molecules that have some sequence identity to each other, but differ in the presence of mutations, is typically used as a substrate for reassembly (&/or one or more additional directed evolution methods described herein). In any given cycle, reassembly (&/or one or more additional directed evolution methods described herein) can occur *in vivo* or *in vitro*, intracellularly or extracellularly. Furthermore, diversity resulting from reassembly (&/or one or more additional directed evolution methods described herein) can be augmented in any cycle by applying prior methods of mutagenesis (e.g., error-prone PCR or cassette mutagenesis) to either the substrates or products of reassembly (&/or one or more additional directed evolution methods described herein). In some instances, a new or improved property or characteristic can be achieved after only a single cycle of *in vivo* or *in vitro* reassembly (&/or one or more